DOCUMENT RESUME

ED 120 265 TM 005 229

AUTHOR O'Reilly, Robert P.; And Others
TITLE Validation of a Multiple-Choice Cloze Test of Literal Comprehension: Summary Report.
PUB DATE Feb 76
NOTE 77p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, Calif., April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage
DESCRIPTORS *Cloze Procedure; Elementary Secondary Education; Factor Analysis; Item Analysis; Item Banks; *Multiple Choice Tests; *Reading Comprehension; *Test Construction; Test Reliability; *Test Validity

ABSTRACT

This report summarizes the development and validation of a test of literal comprehension based on a modified cloze technique. This modification of the cloze offers an objective and partially computerized procedure for constructing items in the multiple-choice format. A field test of the multiple-choice cloze (MCC) involved the administration of 36 test forms across three levels to 5,722 students in grades 1-9. A second measure of literal comprehension based on Bormuth's wh-item was adminstered for comparison. An examination of descriptive data indicates that the MCC and wh-item tests are generally consistent both within themselves and across the two test types. Both tests also proved to be highly reliable with a KR-20 range of .94 to .97 for the MCC, and .85 to .96 for the wh-item test. The construct validity of the MCC was supported by generally appropriate correlations with the wh-item, I.Q., California Achievement Test subscores, and a measure of passage independence. With the exception of one level, the factor analyses generated a first factor interpreted as literal comprehension and defined by the cloze, wh-item, and appropriate CAT subscores. A minor second factor included high loadings on I.Q. and inferential subscores of the CAT. (Author)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED120265

# Validation of a Multiple-Choice Cloze Test of Literal Comprehension: Summary Report*

Robert P. O'Reilly, R. T. Schuder, and Steven J. Kidder

February 1976

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLI

Bureau of School and Cultural Research
New York State Education Department
Albany, New York

TM005 229

*A paper presented at the annual meeting of the National Council on Measurement in Education held in conjunction with the Annual American Educational Research Association, San Francisco, California, April 19-23, 1976.

## Abstract

This report summarizes the development and validation of a test of literal comprehension based on a modified cloze technique. This modification of the cloze offers an objective and partially computerized procedure for constructing items in the multiple-choice format. A field test of the multiple-choice cloze (MCC) involved the administration of 36 test forms across three levels to 5,722 students in grades 1-9. A second measure of literal comprehension based on Bormuth's wh-item was administered for comparison. An examination of descriptive data indicates that the MCC and wh-item tests are generally consistent both within themselves and across the two test types. Both tests also proved to be highly reliable with a KR-20 range of .94 to .97 for the MCC, and .85 to .96 for the wh-item test. The construct validity of the MCC was supported by generally appropriate correlations with the wh-item, I.Q., California Achievement Test subscores, and a measure of passage independence. With the exception of one level, the factor analyses generated a first factor interpreted as literal comprehension and defined by the cloze, wh-item, and appropriate CAT subscores. A minor second factor included high loadings on I.Q. and inferential subscores of the CAT.

More than 20 years of cloze research has neither produced an entirely satisfactory comprehension test nor has it silenced the critics of the cloze technique. While the standard cloze procedure has many technical and theoretical advantages over traditional items for measuring comprehension, it also has some serious liabilities. Most research on the cloze as a measure of comprehension is based on the any-word, every-fifth-word deletion pattern, resulting in a measure that loads too heavily on syntax and general verbal ability (Rankin, 1974). The standard or free-response cloze also has the decided disadvantage of requiring hand scoring and considerable doubt has been cast on the so-called "standard" criteria for scoring and interpreting the test (O'Reilly, Kidder, and Schuder, 1976). In a recent attempt to use the standard cloze as a general alternative to traditional measures of reading comprehension in Madison, Wisconsin (Hansen and Hesse, 1974), serious problems were experienced with the face validity of the technique and with the interpretation and use of the resultant data in the community.

The project reported here modified the standard cloze into a multiple-choice format in an attempt to create a test of reading comprehension that would preserve the obvious advantages of the cloze technique while solving some of its more outstanding theoretical and technical problems. This is a worthwhile goal because the cloze format offers some fundamental improvements for the construction and use of tests of reading comprehension. Cloze tests are highly passage dependent (the student has virtually no chance of responding correctly unless he reads the passage) and a test can be economically constructed in a wholly objective fashion with little test-making expertise. The design of the test format is basically coherent with the ongoing act of comprehension (Smith, 1975) which is characteristically

rapid and focused on the processing of relatively large informational units. The structure and content of the passages in a given cloze test are essentially unchanged by the item format (standard or multiple-choice). The act of responding to the test remains focused on the meaning of the passage, interrupted only by periodic deletions. In this act, the reader predicts the words that correctly replace these deletions to complete the intended meaning by drawing upon his own relevant cognitive structures, unique or otherwise, and not, as is generally the case with other comprehension item formats, those in the mind of a test-maker.

## The Multiple-Choice Cloze Format

The general model for the design of the multiple-choice cloze exercises described here is Hively's (1974) concept of the item form, an objective-generative procedure for producing large numbers of domain-referenced items that are an unbiased representation of the universe of content. The item form, as applied here to the general area of written discourse, consists of an objective or rule-oriented procedure for processing brief, approximately 70-word passages into a 10-item multiple-choice cloze unit. These units were in turn to be ordered in difficulty and assembled into a test or tests of 50 items or more. The objective of this effort was to construct a flexible test-assembly resource that could be used to generate large numbers of tests of literal comprehension that would incorporate unbiased samples of written discourse relevant to the K-12 student population. Such tests could, in turn, be linked with certain domains of written discourse with specified properties of readability, content, etc., thus potentially making reading test scores more directly useful in reading instruction.

Procedures for Clozing Passages

The rules system for initially processing a passage in the multiple-choice cloze format is generally the same as that used for the any-word cloze. However, only nouns, verbs, adjectives, and adverbs are deleted[1] on the assumption that such words carry most of the information which is unique to any given discourse. Function words (determiners, prepositions, auxiliary verbs, etc.) certainly convey information too, but such information is mostly structural; that is, function words primarily define the interrelationships between the appropriate senses of the lexical items of the discourse. Moreover, though syntactic analysis is vital to the comprehension of any verbal message, the information communicated by syntactic analysis alone is not unique to a particular message or content area. Syntax rarely tells the reader anything he does not already know (Katz, 1972).

Research on the cloze also suggests that the deletion of nouns, verbs, and modifiers will tend to reduce correlations between cloze scores and measures of general verbal ability or IQ while simultaneously increasing correlations with measures of "specific comprehension" and "information gain" (Rankin 1974; Bormuth, 1971). Substantial or high correlations of IQ with multiple-choice cloze test scores were considered inappropriate in view of the fact that the test was conceptualized as a measure of the comprehension of the literal meaning of written discourse.

Fillenbaum et al. (1963) found that the grammatical and semantic information necessary to replace nouns, verbs, and modifiers in written discourse tended to be further removed from the deletion than the informa-

[1]Only nouns and verbs are deleted in grade 1 and 2 materials.

tion needed to replace function words. Topical content, for instance, which is more likely to affect the particular choice of nouns, verbs, or modifiers in a given text, is liable to be dispersed throughout the text. It is assumed, therefore, that the restriction of the deletion type to nouns, verbs, and modifiers will obviate some of the criticism of the cloze as a measure of local redundancy and increase the semantic component of the text in cloze scores without eliminating the syntactic component. Nouns, verbs, and modifiers also convey structural information, albeit to a lesser degree than function words.

Finally, deleting only nouns, verbs, adjectives, and adverbs makes it feasible to construct a cloze test in the multiple-choice format. That is, relatively few words can function as determiners or auxiliary verbs in an English sentence; making up distractors for "the," for instance, seems pointless. On the other hand, the number of words that can function as nouns in an English sentence is enormous.

Deletion Rate

An every-fifth-word deletion rate (20% of the text) is considered optimum because it samples the grammatical and semantic relations of the text objectively and as thoroughly as possible without depriving the student of the information he needs to replace the deleted words. Selective deletion types (e.g., nouns), however, force some variation in deletion rate. For example, the test-maker who is trying to maintain an every-fifth-word deletion rate while deleting only nouns, verbs, and modifiers, often counts five words in a text and finds no candidate for deletion. If he backs up below three words between deletions, it becomes very difficult to replace missing words with so little remaining, immediate context. On the other hand, if he counts too far in the other direction,

then the thoroughness of the sample of the grammatical and semantic relations begins to suffer. A test-maker cannot readily reject passages on the grounds that they create difficulties for the adopted deletion pattern, else the passages will represent a biased selection from the domain of reading materials. Consequently, the every-fifth-word deletion rate is adhered to as much as possible; there are never fewer than three words between deletions; occasionally there are as many as eleven words between deletions.

Distractors

All distractors are grammatically plausible and semantically implausible. Grammatical plausibility means that any distractor can perform properly the grammatical function assigned to it by the syntactic position of the missing word for which it functions as a distractor. For example, if a noun is deleted, the distractors are "nouns" or words that can behave like nouns in the syntactic position in question. Distractors are grammatically plausible in order to reduce further the possibility that syntactic information alone will be used to determine the correct answer to a given deletion.

Since the distractors are all grammatically plausible, syntactic analysis provides no definitive basis for choosing among the sets of responses. On the other hand, syntactic analysis is fundamental to the comprehension of the interrelationships among the words of any sentence, and no test of comprehension should or could short-circuit that analysis. Consequently, the sets of responses associated with each deletion are located below each passage, encouraging the student to process the grammatical and semantic information in the text, to make an initial hypothesis about the meaning and identify of the missing word before looking for

verification of that hypothesis among the sets of responses. In other words, an attempt is made to preserve the best feature of the standard cloze procedure--its similarity to the reading process--while introducing a multiple-choice format. In summary, syntax is present in the modified cloze format adopted in the multiple-choice cloze passages, but it is not a factor in choosing among the distractors.

In the set of responses associated with any deletion, only the correct answer, the exact word in the original text, is semantically plausible. Again, the attempt is made to preserve the best features of the standard cloze procedure while weeding out its liabilities. In the standard cloze procedure, the test is scored for the exact word supplied by the author of the passage, a procedure which seems antithetical to models of reading as a constructive language process. On the other hand, in that search for the basis of the commonality of meaning intended by the writer and apprehended by the reader which is literal comprehension, it was posited that the common meeting ground between writers and readers is the orthographic representation of meaning on the printed page, and therefore that representation is the closest possible approximation of what the writer meant. Accordingly, the modified cloze procedure maintains the exact-word-only replacement criterion. The student who makes a correct hypothesis about the meaning of a missing word based upon his apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse" (Katz and Fodor 1963, p. 172) will have no difficulty modifying his hypothesis about the surface representation of that meaning when confronted with the correct answer among a set of distractors. The distractors only behave like traditional distractors when the vocabulary level or the syntactic complexity of the passage exceeds the student's competence, that is, when

the student can no longer comprehend at the literal level.

No attempt is made to tamper with the approximation of the meaning intended by the author. No attempt is made to interpret the text for the student, to impose the test-writer's own idiosyncratic interpretation of the text on the text in the form of distractors that compete with the correct response. Such semantic competition, such alternate possibilities for interpretation, are inappropriate to the underlying construct for the test, literal comprehension, which is rooted in commonality of interpretation rather than nuances of meaning. Cranney (1972) found, for instance, that semantically plausible distractors in a multiple-choice cloze format introduce significant numbers of items into the cloze test that are even more difficult than the hardest items on standard cloze tests. Semantically plausible distractors extend the context for interpretation beyond the grammatical and semantic relations of the discourse which is, again, antithetical to the construct, literal comprehension. It is hypothesized that semantically plausible distractors will also maintain or increase correlations between the cloze procedure and measures of general verbal ability or IQ. Finally, semantically plausible distractors, with an emphasis on nuances of meaning, make the cloze procedure into a very difficult vocabulary test. Literal comprehension does not demand so extensive or refined a vocabulary.

## The Construction of the Multiple-Choice Cloze Exercises

The construction of the multiple-choice cloze exercises was undertaken in order to test the efficacy of the multiple-choice cloze format as a measure of literal comprehension and as a means of readily implementing, with reproducible test items, the concept of domain-referenced testing. At the outset, a plan was devised for the systematic sampling of reading

materials in four domains in which students are expected or required to read. The domains are:

1. Textual Material in Reading/Literature, Language Arts, Social Studies, Science and Mathematics;
2. Citizen Material (newspapers and news magazines);
3. Consumer Material (catalogs, advertising, instructions, and so forth); and
4. Reference Material (test instructions, children's magazines, encyclopedias, and so forth).

Selection of Cloze Passages

Textual materials were to be sampled at each grade level, from 1 through 10. Materials in the other domains were to be assigned to grade levels on the basis of readability scores. Quotas were established for the number of samples to be collected, at random, for each grade and domain. The resources used for the sample collection were the New York State Education Department's Curriculum Laboratory and the State Library. In addition, some consumer passages were taken from A Pilot Reading Literacy Assessment of Madison Public School Students (Hansen and [illegible]sse, 1974).

The selection procedures resulted in the identificat[illegible] of 1,374[2] passages that were coherent and of specified lengths a[illegible]e for clozing. Their distribution by content domain and grad[illegible]ability is shown in Table 1. Table 1 also shows the distribution of the textual materials by subject matter.

Determination of Readability

All of the passages--those in the textual domain as well as those in the citizen, consumer, and reference domains--were subjected to readability

[2]Another 120 passages have been added to the citizen domain and expansion of the textual domain into college levels is anticipated.

Table 1

Distribution of Cloze Passages by Graded Readability Level and Content Area

| Read-ability Level | Domain: Textual | | | | | | Citizen | | Consumer | Reference | Grand Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reading | Lang. Arts | Math | Sci. | Soc. St. | Total | Newsp. | Mag. | | | |
| 1 | 11 | | | | 1 | 12 | | | | | 12 |
| 2 | 30 | | | | 10 | 40 | | | | | 40 |
| 3 | 21 | 1 | | | 14 | 36 | | | | | 36 |
| 4 | 26 | 0 | | | 26 | 52 | | | | | 52 |
| 5 | 13 | 12 | 7 | 12 | 22 | 66 | | | | 4 | 70 |
| 6 | 16 | 9 | 8 | 4 | 9 | 46 | | | | 6 | 52 |
| 7 | 26 | 6 | 4 | | 7 | 43 | | | | 5 | 48 |
| 8 | 13 | 12 | 9 | 3 | 8 | 45 | | | | 4 | 49 |
| 9 | 28 | 8 | 3 | 3 | 7 | 49 | 1 | | | 4 | 54 |
| 10 | 14 | 10 | 8 | 6 | 14 | 52 | 2 | | | 6 | 60 |
| 11 | 19 | 9 | 8 | 3 | 5 | 44 | 2 | | 5 | 6 | 57 |
| 12 | 10 | 14 | 7 | 9 | 11 | 51 | 4 | | 3 | 6 | 64 |
| 13 | 19 | 12 | 9 | 8 | 14 | 62 | 2 | 3 | 3 | 5 | 75 |
| 14 | 7 | 11 | 10 | 11 | 13 | 52 | 4 | 3 | 6 | 6 | 71 |
| 15 | 19 | 5 | 9 | 8 | 12 | 53 | 3 | 7 | 4 | 5 | 72 |
| 16 | 16 | 11 | 10 | 13 | 12 | 62 | 2 | 6 | 5 | 5 | 80 |
| 17 | 16 | 9 | 9 | 17 | 6 | 57 | 10 | 12 | 7 | 5 | 91 |
| 18 | 19 | 10 | 9 | 11 | 9 | 58 | 6 | 7 | 8 | 5 | 84 |
| 19 | 8 | 9 | 8 | 6 | 5 | 36 | 7 | 8 | 5 | 5 | 61 |
| 20 | 12 | 6 | 16 | 11 | 7 | 52 | 13 | 9 | 6 | 5 | 85 |
| 21 | 3 | 3 | 7 | 5 | 3 | 21 | 13 | 10 | 7 | 4 | 55 |
| 22 | 7 | 3 | 8 | 6 | 2 | 26 | 7 | 7 | 6 | 4 | 50 |
| 23 | | | 6 | 7 | 3 | 16 | 10 | 11 | 6 | 5 | 48 |
| 24 | | | 4 | 5 | | 9 | 9 | 8 | 7 | 5 | 38 |
| 25 | | | | 5 | | 5 | 7 | 10 | 7 | | 29 |
| 26 | | | 1 | 6 | | 7 | 7 | 6 | 5 | | 25 |
| 27 | | | | 1 | | 1 | 7 | 8 | 4 | | 20 |
| 28 | | | | | | | 4 | 5 | 6 | | 15 |
| Total | 353 | 160 | 160 | 160 | 220 | 1,053 | 120 | 120 | 100 | 100 | 1,493 |

calculations so that they could be eventually ordered by difficulty for test construction purposes. The readability formulas used were the Spache (1953, 1960) and the Dale-Chall (1948). The range of Spache scores was divided into six equal intervals, and the range of Dale-Chall scores was divided into 22 equal intervals. This gave 28 difficulty levels covering grades 1 through college. The raw scores, difficulty levels, and original grade level interpretations given by Dale-Chall and Spache are shown in Table 2.

Use of the readability formulas disclosed wide ranges of difficulty among instructional materials at given grade levels. Both extremely easy and extremely difficult passages that differed markedly from other materials for the same grade were eliminated in the selection process. However, there is still variation in the number of difficulty levels (see Table 1) covered by grade levels in the textual and other domains due to differences in the distributions of readability in a given domain of content. The grade level of the source was used as the guide in assigning a cloze passage to a grade level. Both grade level and difficulty level are indicated by the identification number for each passage.

Preparation of Cloze Items

The procedure for word deletion in the cloze passages varied with the grade of the source. In grade 1 and 2 materials, every eighth word was deleted, and deletions were limited to nouns and verbs. For grade 3 and above, every fifth word was deleted. Deletions included adjectives and adverbs as well as nouns and verbs.

In all cases, the initial deletion was made between the sixth and tenth words. The exact starting point was determined by a table of random numbers. The number of deletions per passage was fixed by the passage

Table 2

Difficulty Levels for Cloze Passages

| Readability formula | Raw score | Difficulty level | Original grade level assignments by Spache and Dale-Chall |
|---|---|---|---|
| S | 1.0-1.4 | 1 | 1 |
| P | 1.5-1.9 | 2 | |
| A | 2.0-2.4 | 3 | 2 |
| C | 2.5-2.9 | 4 | |
| H | 3.0-3.4 | 5 | 3 |
| E | 3.5-3.9 | 6 | |
| D | 4.50-4.74 | 7 | 4 |
| A | 4.75-5.99 | 8 | |
| L | 5.00-5.24 | 9 | 5-6 |
| E | 5.25-5.49 | 10 | |
| | 5.50-5.74 | 11 | |
| C | 5.75-5.99 | 12 | |
| H | 6.00-6.24 | 13 | 7-8 |
| A | 6.25-6.49 | 14 | |
| L | 6.50-6.74 | 15 | |
| L | 6.75-6.99 | 16 | |
| | 7.00-7.24 | 17 | 9-10 |
| | 7.25-7.49 | 18 | |
| | 7.50-7.74 | 19 | |
| | 7.75-7.99 | 20 | |
| | 8.00-8.24 | 21 | 11-12 |
| | 8.25-8.49 | 22 | |
| | 8.50-8.74 | 23 | |
| | 8.75-8.99 | 24 | |
| | 9.00-9.24 | 25 | 13-15 (College) |
| | 9.25-9.49 | 26 | |
| | 9.50-9.74 | 27 | |
| | 9.75-9.99 | 28 | |

length, which varied by grade level. The number of alternatives in the multiple-choice responses also varied by grade level: three alternatives at grade 1, four at grades 2 and 3, and five at grades 4 and above. These variations by grade level are summarized in Table 3, Specifications for Cloze Passages and Test Items.

The correct multiple-choice response to a cloze item is the exact word deleted from the passage. To assure distractors of appropriate difficulty for the test items, graded lists of nouns, verbs, adjectives, and adverbs were prepared using Harris and Jacobson's Basic Elementary Reading Vocabularies (1972) and EDL Research and Information Bulletin 5: A Revised Core Vocabulary (Taylor, Frackenpohl, and White, 1969). Special content words for subject matter areas like Social Studies were compiled using the Harris-Jacobson material and the American Heritage Word Frequency Book (Carroll, Davies, and Richman, 1971).

Initially, distractors were selected from appropriate lists by use of a table of random numbers. Later, a computer program was written for automatic random selection of distractors. Each set of distractors was reviewed to eliminate tricky or ineffective distractors, such as synonyms, and to assure that the distractors agreed with the stem in tense, number, and so forth.

With a minimum of 3 deletions per passage at grade 1 and a maximum of 10 deletions per passage at grade 3 and above, nearly 15,000 multiple-choice items have been prepared for the cloze exercises.

Format

All cloze passages and test items were put in a comparable format. The format gives (1) the identification number of the passage, (2) a title (provided by the item writer), (3) the passage itself, and (4) the test

Table 3

Specifications for Cloze Passages and Test Items

| | Grade 1 | Grade 2 | Grade 3 | Grade 4 and above |
|---|---|---|---|---|
| Passage length | 25-35 words | 40-45 words | 60-70 words | 60-70 words |
| Words deleted | Nouns Verbs | Nouns Verbs | Nouns Adjectives Verbs Adverbs | Nouns Adjectives Verbs Adverbs |
| Frequency of deletions | Every 8th word | Every 8th word | Every 5th word | Every 5th word |
| Deletions per passage | 3 | 5 | 10 | 10 |
| Alternatives per item | 3 | 4 | 4 | 5 |

## WHAT DOES ANDY SEE?

Andy saw something at his bedroom
____1____! He ran as fast as he could to
____2____ Mother.
"Listen, Mother!" said Andy.
Mother said, "Please ____3____, Andy. I have to get ready to ____4____ to work now. Mrs. Coats is ____5____ for me."

(1)
a. paper
b. window
c. apple
d. oven

(2)
a. wash
b. tell
c. buy
d. fix

(3)
a. peep
b. fly
c. point
d. wait

(4)
a. think
b. sing
c. tip
d. go

(5)
a. finding
b. roping
c. waiting
d. racing

Figure 1. Sample cloze passage and items.

items. Large (Bulletin) type was used for the first two grades. A sample cloze passage for grade 2 is shown in Figure 1.

## Organization of the Research

The first field test of the multiple-choice cloze testing materials was designed to fulfill several purposes: (1) It explored the use of the testing materials in applying a survey design in one textual area; (2) it provided data for detailed item analyses; (3) it provided an initial test of the ability of the testing system in assembling large numbers of parallel test forms; (4) it provided a basis for testing out the Rasch or latent-trait model as an approach to scaling; (5) it made available reliability data on a large number of test forms; and (6) it provided initial evidence on the validity of the construct. Reported here are the findings relating to (3), (5), and (6). A full report of the research findings is presented in O'Reilly et. al. (1976).

The design for data collection required two 40-minute testing periods, one to collect response data on the multiple-choice cloze items, and one for the collection of response data on an alternate measure of the construct of literal comprehension developed for the study. In addition, standardized achievement scores in reading and language arts were collected by the study district at the time of the two experimental test administrations. The test administration sample is given in Table 4, identified in relation to the two experimental test administrations, one of which involved the multiple-choice cloze and the other of which involved a test called the wh-item test. Each student in the study sample had some advance training in the cloze format. This ranged from several 20-minute training sessions in the primary grades to a 15-minute training session at grades 4 and above. The characteristics of the various tests used in the study and the conditions of their administration follow:

Table 4

Number of Students Tested per Form, Multiple-Choice Cloze Test and Wh-Item Test

| Multiple-Choice Cloze Test (N = 5,264)[a] | | | | | | Wh-Item Test (N = 5,197)[a] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Level I | | Level II | | Level III | | Level I | | Level II | | Level III | |
| Form | N | Form | N | Form | N | Form | N | Form | N | Form | N |
| 1 | 128 | 13 | 147 | 25 | 167 | 37 | 127 | 49 | 147 | 51 | 163 |
| 2 | 126 | 14 | 152 | 26 | 164 | 38 | 124 | 50 | 152 | 62 | 162 |
| 3 | 130 | 15 | 153 | 27 | 160 | 39 | 126 | 51 | 148 | 63 | 164 |
| 4 | 124 | 16 | 152 | 28 | 161 | 40 | 121 | 52 | 152 | 64 | 161 |
| 5 | 126 | 17 | 146 | 29 | 158 | 41 | 119 | 53 | 145 | 65 | 165 |
| 6 | 126 | 18 | 151 | 30 | 165 | 42 | 122 | 54 | 144 | 66 | 166 |
| 7 | 127 | 19 | 152 | 31 | 158 | 43 | 124 | 55 | 145 | 67 | 154 |
| 8 | 127 | 20 | 148 | 32 | 163 | 44 | 124 | 56 | 149 | 68 | 163 |
| 9 | 129 | 21 | 152 | 33 | 166 | 45 | 131 | 57 | 147 | 69 | 164 |
| 10 | 127 | 22 | 152 | 34 | 159 | 46 | 123 | 58 | 148 | 70 | 156 |
| 11 | 120 | 23 | 148 | 35 | 163 | 47 | 121 | 59 | 157 | 71 | 163 |
| 12 | 123 | 24 | 149 | 36 | 165 | 48 | 121 | 60 | 145 | 72 | 154 |
| All | 1,513 | All | 1,802 | All | 1,949 | All | 1,483 | All | 1,779 | All | 1,935 |

Note. Total N = 5,722. Distribution by grade: 1, 551; 2, 536; 3, 524; 4, 580; 5, 649; 6, 692; 7, 751; 8, 709; 9, 730.

[a]Not all students were present for all tests.

Multiple-Choice Cloze Test Forms

The multiple-choice cloze test forms consisted of three sets of 12 parallel test forms designated Level I for grades 1-3, Level II for grades 4-6, and Level III for grades 7-9. The test forms were assembled as depicted in Table 5. Each test level was assigned a range of readability in the cloze passage pool and within this range of readability consecutive pairs of readability levels were regarded as sampling units (except at Level I where the first two readability levels were treated as separate sampling units). Test forms in a given test level were then constructed by sampling without replacement 6 ordered passages, one from each consecutive sampling unit. The order of passage readability was maintained in the test form. Each test form contained from 39 to 60 items presented to the testee in groups of 3 to 10 items per passage depending on the test level.

Wh-Item Test

Because the CAT was, in many respects, an ambiguous criterion for the multiple-choice cloze as a measure of literal comprehension, an alternative measure of the construct was developed. This test, called the wh-item test was assembled from an original pool of some 300 passages and 3,000 main idea and wh-items using a design for the selection of test passages that was virtually identical to that used for the assembly of the cloze test forms. Five wh-items were then selected from each of the 8 wh-items available on a passage in a manner that resulted in equal distributions of wh-item types in a test form in a test level. The wh-item types are: how, what [noun], what [verb], when, where, which, who, why. This procedure resulted in a set of uniform 30-item tests in each test level that paralleled the multiple-choice cloze tests in number of passages and range of passage difficulty.

Table 5

Sampling Design for Survey Test in the Textual Domain--
Reading/Literature for Cloze Exercises

| Sampling unit: Readability level | Passage Pool by Grade Level of Source | | | | | | | | | | Passages per Sampling Unit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| 1 | 11 | | | | | | | | | | 1 |
| 2 | 30 | Level I | | | | | | | | | 1 |
| 3 | 7 | 14 | | | | | | | | | 1 |
| 4 | | 26 | | | | | | | | | |
| 5 | | | 12 | 1 | | | | | | | 1 |
| 6 | | | 12 | 4 | Level II | | | | | | |
| 7 | | | 6 | 13 | 7 | | | | | | 1 |
| 8 | | | | 9 | 4 | | | | | | |
| 9 | | | | 11 | 11 | 6 | | | | | 1 |
| 10 | | | | 4 | 6 | 4 | | | | | |
| 11 | | | | | 6 | 11 | 2 | | | | 1 |
| 12 | | | | | 2 | 5 | 3 | Level III | | | |
| 13 | | | | | | 6 | 11 | 2 | | | 1 |
| 14 | | | | | | 1 | 5 | 1 | | | |
| 15 | | | | | | | 5 | 10 | 4 | | 1 |
| 16 | | | | | | | 4 | 8 | 4 | | |
| 17 | | | | | | | | 6 | 5 | 5 | 1 |
| 18 | | | | | | | | 3 | 9 | 7 | |
| 19 | | | | | | | | | 4 | 4 | 1 |
| 20 | | | | | | | | | 4 | 8 | |
| 21 | | | | | | | | | | 3 | 1 |
| 22 | | | | | | | | | | 7 | |

Test-Wiseness Test

Because multiple-choice tests of reading comprehension are vulnerable to a form of test-wiseness referred to as "passage independence," a special test was constructed to measure some aspects of this characteristic. Referred to as the test-wiseness test, the design of these tests paralleled the cloze and wh-item test form designs. The wh-items in each of the three test levels were pooled and systematically assigned in units of 12 to each wh- test form, such that no test items on this part of a given test were referenced to the passages on that test. Care was also taken to represent the passage difficulties and types of wh-items in a test level in an attempt to create parallel test forms. The relationship between scores on this test-wiseness measure and scores on the wh-item test provides some indication of the extent to which students' responses on the latter test are dependent on reading the associated test passages. This test also provides some indication of the extent to which this form of test-wiseness affects responses on the multiple-choice cloze test. However, it is probable that the cloze test is vulnerable to other forms of test-wiseness, a possibility that should be investigated directly.

Short Form Test of Academic Aptitude

The Short Form Test of Academic Aptitude (SFTAA) is a group-administered intelligence test that yields language and non-language IQ's. This test, administered by the school district to students in grades 1-6, along with the California Achievement Test, permitted study of the relationship between IQ and the literal comprehension tests across the study subsample.

California Achievement Test

The California Achievement Test (CAT), Form A, was administered to students in grades 1-8. The CAT has four levels as follows:

| CAT Level | Grade(s) |
|---|---|
| I | 1 |
| II | 2-3 |
| III | 4-6 |
| IV | 7-8 |

The CAT skill (major headings) and subtest (minor headings) scores used in the study are identified in Table 6 by test level, along with the numbers of items in each skill and subtest score.

## Analysis

The data set available on the foregoing test scores was organized for analysis by CAT level and by wh-item and multiple-choice cloze test level. There were four such subgroups: (a) students in grade 1 who took Level I on the CAT and Level I on the wh-item and multiple-choice cloze tests (N = 456); (b) students in grades 2 and 3 who took Level II on the CAT and Level I on the wh-item and multiple-choice cloze tests (N = 972); (c) students in grades 4, 5, and 6 who took Level III on the CAT and Level II on the wh-item and multiple-choice cloze tests (N = 1,697); and (d) students in grades 7 and 8 who took Level IV on the CAT and Level III on the wh-item and multiple-choice cloze tests (N = 1,260).

To permit an analysis across test forms at each cloze test level, the raw scores for the wh-item and multiple-choice cloze test forms were converted to $z$ scores based on the score distribution for each test form in a test level. Subsequently, negative values were eliminated by applying a linear transformation to each set of obtained $z$ scores. The resultant scores from any of the 12 wh-item and multiple-choice cloze tests in a test level were thereafter treated as having come from equivalent test forms and were combined as required for the analyses by CAT level. This approach to test equating, though somewhat unorthodox, is defensible on several grounds. The general shapes, means, and the standard deviations of the

TABLE 6

Number of Items Per CAT Skill and Subtest Category by Test Level

| Skill and Subtest Category | CAT Level | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| Reading Vocabulary | 92 | 40 | 40 | 40 |
| Sentence-Picture Association | 10 | | | |
| Beginning Sounds | 10 | | | |
| Ending Sounds | 10 | | | |
| Letter Recognition | 15 | | | |
| Word Form | 10 | | | |
| Picture-Word Association | 10 | | | |
| Word Recognition | 12 | 20 | | |
| Words in Context | 15 | 20 | 40 | 40 |
| Reading Comprehension | 24 | 45 | 42 | 45 |
| Facts | 7 | 12 | 11 | 5 |
| Interpretation | 6 | 9 | 15 | 11 |
| Relationships | | | 4 | 7 |
| Generalizations | | 6 | 9 | 11 |
| Inferences | 8 | 8 | 4 | 11 |
| Content-General[1] | | | 11 | 10 |
| Content-Social Studies | | | 7 | 10 |
| Content-Science | | | 10 | 10 |
| Content-Mathematics | | | 8 | 9 |
| Language Mechanics | 37 | 66 | 80 | 72 |
| Language Usage | 20 | 25 | 41 | 50 |
| Standard English | 20 | 25 | 24 | 28 |
| Sentence Structure | | | 5 | 6 |
| Sentence Parts and Functions | | | 7 | 10 |
| Transformations | | | 5 | 6 |

[a]The items used in the content area categories were also included in the other comprehension categories. Hence, there is an overlapping of items in the content areas with items in the other comprehension categories in Levels III and IV.

distributions of wh-item and multiple-choice cloze test scores were very similar from form to form in a test level (usually the average raw score difference from form to form was less than one-fourth of a standard deviation), the internal reliabilities of each form were consistently high, and the tests had been systematically assembled to be parallel in order and range of readability level.[3]

The comparability of the cloze and wh-item tests as measures of a common construct of literal comprehension was initially examined by intercorrelating the various subscores and total scores on these tests by test level. The test-wiseness measure was included in these analyses. The reliability of all three types of tests was estimated by applying the Kuder-Richardson Formula 20 to the 108 available test forms.

The issue of construct validity is first approached in an examination of the zero-order correlations of the multiple-choice cloze test with the other test criteria available in the study. A priori specifications are made about the expected correlations between the multiple-choice cloze test and the wh-item test with all of the skill and subscores derived from the CAT. When expected correlations are not obtained, a detailed study is made of the possible causes for the unexpected results. These detailed studies focus directly on analyzing the item content of the CAT subtest in question in relation to the construct of literal comprehension.

In a final examination of construct validity, the results of a factor analysis run on the major test scores in the study is presented for each level of the sample. These analyses provide an additional exploration of the validity of the construct of literal comprehension which is expected

---

[3]This same procedure was followed for the test-wiseness test forms, but here as will be seen the justification for equivalence of test forms is less adequate.

to be resolved as a reading comprehension factor distinct from the traditional reasoning or inferential factor of comprehension.

To provide a basis for interpreting the results that follow, Table 7 displays by CAT level the means and standard deviations of scores on the major variables under consideration. It will be observed that the multiple-choice cloze, wh-item, and test-wiseness tests were standardized (not normalized) within CAT level. The major subtests on the CAT are normalized. The derived subscores on the CAT are raw scores.

There are several characteristics of this data set that must be taken into consideration when interpretations are made of the zero-order, Pearson product-moment correlations between variables. The construct validity sections of this report are based entirely on these correlations.

It is well-known that the reliability of a test as well as the shape of its score distribution affects its correlation with another test or measure. Reliability, as shown in the next section, does not contribute greatly to ambiguity in interpreting the results reported here. The wh-item and cloze tests meet the most conservative standards of reliability. The test-wiseness test is sufficiently reliable for the purpose used here. A few of the CAT subtests previously listed in Table 6 should be suspect because of length; they are, at Level III, Relationships (4 items), Inferences (4 items), Sentence Structure (5 items) and Transformations (5 items), and, at Level IV, Facts (5 items). The subscores for the wh-item tests, which are based on small numbers of each of the wh-item types in the test, and the deletion type subscores for adjectives and adverbs on the cloze test, which are typically based on very few items, undoubtedly contribute to reduced correlations among the subscores and total scores for these tests. However, these relationships were as expected, and they

Table 7

Means and Standard Deviations of Scores on the Multiple-Choice Cloze, Wh-Item, Test-Wiseness, California Achievement, and Short Form Test of Academic Aptitude Tests

| Test | CAT level I (N=456) | | II (N=972) | | III (N=1,697) | | IV (N=1260) | |
|---|---|---|---|---|---|---|---|---|
| | $\bar{X}$ | SD | $\bar{X}$ | SD | $\bar{X}$ | SD | $\bar{X}$ | SD |
| Multiple-Choice Cloze Test[a] | 49.58 | 9.97 | 49.58 | 9.89 | 49.80 | 9.82 | 49.64 | 9.85 |
| Wh-Item Test | 49.59 | 9.94 | 49.63 | 9.83 | 49.71 | 9.90 | 49.65 | 9.84 |
| Test-Wiseness Test | 49.71 | 9.79 | 49.53 | 9.86 | 49.61 | 9.91 | 49.59 | 9.95 |
| California Achievement Test | | | | | | | | |
| Reading Vocabulary-ADSS[b] | 316.11 | 34.52 | 370.14 | 46.78 | 430.95 | 64.63 | 502.41 | 92.90 |
| Reading Comprehension-ADSS | 298.39 | 48.29 | 389.00 | 65.03 | 461.12 | 68.67 | 518.18 | 91.47 |
| Language Mechanics-ADSS | 305.51 | 53.11 | 386.65 | 68.76 | 469.05 | 87.44 | 529.53 | 120.99 |
| Language Usage-ADSS | 327.08 | 57.48 | 393.22 | 67.97 | 459.21 | 75.06 | 511.57 | 107.59 |
| Language IQ--SFTAA[c] | 102.09 | 13.07 | 100.54 | 14.61 | 98.71 | 14.21 | - | |
| Non-Language IQ--SFTAA | 104.31 | 15.6[illegible] | 104.15 | 14.90 | 104.32 | 14.87 | - | |
| Vocabulary Subscores[d] | | | | | | | | |
| Sentence-Picture Association | 9.87 | .45 | - | | - | | - | |
| Beginning Sounds | 8.02 | 1.82 | - | | - | | - | |
| Ending Sounds | 8.96 | 1.43 | - | | - | | - | |
| Letter Recognition | 14.54 | 1.57 | - | | - | | - | |
| Word Form | 8.41 | 2.09 | - | | - | | - | |
| Picture-Word Association | 7.19 | 2.42 | - | | - | | - | |
| Word Recognition | 10.04 | 2.47 | 18.50 | 2.49 | - | | - | |
| Words in Context | 5.45 | 3.90 | 14.42 | 4.84 | 24.83 | 8.82 | 24.70 | 8.39 |

Table 7 (Continued)

| Test | CAT level I (N=456) | | II (N=972) | | III (N=1,697) | | IV (N=1260) | |
|---|---|---|---|---|---|---|---|---|
| | $\bar{X}$ | SD | $\bar{X}$ | SD | $\bar{X}$ | SD | $\bar{X}$ | SD |
| Comprehension Subscores: | | | | | | | | |
| Facts | 3.29 | 2.06 | 8.61 | 3.34 | 8.52 | 2.31 | 3.43 | 1.19 |
| Interpretation | 2.12 | 1.54 | 5.69 | 2.62 | 9.76 | 3.49 | 5.92 | 2.81 |
| Relationships | - | | - | | 1.52 | 1.16 | 3.50 | 1.63 |
| Generalizations | - | | 3.87 | 1.82 | 3.97 | 1.94 | 4.19 | 2.15 |
| Inferences | 2.29 | 1.74 | 5.19 | 2.34 | 1.85 | 1.13 | 7.10 | 2.32 |
| Reading-General | - | | - | | 6.24 | 2.60 | 6.12 | 2.43 |
| -Social Studies | - | | - | | 4.59 | 1.71 | 5.42 | 2.40 |
| -Science | - | | - | | 5.08 | 2.32 | 4.46 | 2.32 |
| -Mathematics | - | | - | | 4.02 | 2.17 | 3.43 | 2.18 |
| Language | 11.70 | 3.88 | 14.70 | 5.18 | 14.42 | 3.52 | 15.38 | 3.65 |
| Standard English | - | | - | | 3.35 | 1.22 | 2.34 | 1.56 |
| Sentence Structure | - | | - | | 1.76 | 1.65 | 3.58 | 2.78 |
| Sentence Parts and Function | - | | - | | 2.51 | 1.29 | 3.91 | 1.30 |
| Transformations | | | | | | | | |

[a] Scores on the Multiple-Choice Cloze Test, Wh-Item Test and the measure of Test-Wiseness were standardized according to CAT level for this analysis to have a mean of 50 and SD of 10.

[b] Achievement Development Scale Scores (ADSS) were derived by CTB/McGraw-Hill from a single equal-interval score scale across all grades for use with all levels and forms of the CAT.

[c] IQ scores obtained from Short Form Test of Academic Aptitude given with CAT in grades 1-6.

[d] All California Achievement Test Subscores are raw scores.

do not seem to have contributed to interpretive problems, particularly when the high degree of internal consistency of the wh-item and cloze tests is shown later with other methods.

A minor problem for interpretation of the correlational results derives from the test score distributions of the cloze and wh-item tests which are generally skewed in the negative direction. This skewness is the result of generally easy tests which interact with grade level in this case to produce more extreme negative skewness at the upper grades of a test level and also across the grades of the total study sample. The general result of this skewness is a frequent lack of homoscedasticity at the upper score levels in the bivariate plots of the test score distributions in the present study. This effect appears to have produced attenuation in certain of the correlations at certain levels of the study sample which is predictable as follows.

In grades 1, 2, and 3, and particularly in grade 1, the wh-item and cloze tests were considerable more difficult for students than in other grades. As a result there was no artificial ceiling on the tests, and students' scores were spread more evenly over the score ranges for the two tests. Given the high reliability of the Level I wh-item and cloze tests for grades 1, 2, and 3, the correlations of the Level I cloze and wh-item tests with each other and with the CAT Level I and II skill and subtest scores should be close to the "true" correlations among these variables in the sample.

In CAT Levels III and IV of the study sample, the degree of negative skewness on the wh-item and cloze tests is more marked, with the Level IV data being the most affected. At these test levels, the attenuating effects of skewness on the correlations involving the wh-item and cloze

tests should be noticeable, particularly at Level IV. The overall effect of the skewed distributions of the scores on the two measures of literal comprehension constructed for the study is to reduce the validity coefficients between the cloze and wh-item tests and between these tests and the various CAT scores in the upper levels of the study sample.

A final complicating factor for interpreting the correlations in the study derives from the restricted range of talent in some CAT levels. CAT Level I includes only grade 1 and CAT Levels II and IV include only a two-grade spread. The remaining, CAT III, subsample has a three-grade spread. The narrow range of talent for some CAT levels, however, appears to have restricted the correlations only in CAT Level IV subsample.

## Test Comparability and Reliability

The correlations among subscores and total scores within and between the three types of tests specially constructed for this research are presented in Tables 8, 9, and 10. The three total test scores are based on the wh-item test, the multiple-choice cloze test, and the test-wiseness test. The subscores for the multiple-choice cloze test are each based on one of the four grammatical parts of speech deleted in a passage. For example, each test form has a subscore for the sum of all correct noun responses. The subscores for the wh-test are based on the various types of wh-items included in the test. The test-wiseness test has only a total score.

Tables 11 and 12 present, respectively, the means and standard deviations and the Kider-Richardson Formula 20 reliability coefficients for each of the 36 forms of the three types of tests. All Tables are

Table 8

Intercorrelations of Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests: Total Scores and Subscores
Level I; Grades 1, 2, and 3
(N = 1,445)

| Scores | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCC Test | | | | | | | | | | | | | | | |
| 1 Noun | - | | | | | | | | | | | | | | |
| 2 Verb | .85 | - | | | | | | | | | | | | | |
| 3 Adjective | .70 | .69 | - | | | | | | | | | | | | |
| 4 Adverb | .49 | .45 | .35 | - | | | | | | | | | | | |
| 5 Total | .96 | .94 | .78 | .53 | - | | | | | | | | | | |
| Wh-item Test | | | | | | | | | | | | | | | |
| 6 How | .63 | .62 | .47 | .29 | .64 | - | | | | | | | | | |
| 7 What (N-P) | .64 | .62 | .46 | .32 | .65 | .58 | - | | | | | | | | |
| 8 What (V) | .64 | .61 | .47 | .32 | .65 | .60 | .59 | - | | | | | | | |
| 9 When | .56 | .55 | .41 | .28 | .57 | .50 | .46 | .49 | - | | | | | | |
| 10 Where | .59 | .58 | .43 | .28 | .61 | .54 | .56 | .55 | .50 | - | | | | | |
| 11 Which | .65 | .61 | .47 | .33 | .65 | .63 | .60 | .61 | .51 | .57 | - | | | | |
| 12 Who | .66 | .64 | .46 | .30 | .67 | .63 | .64 | .62 | .52 | .59 | .65 | - | | | |
| 13 Why | .59 | .59 | .44 | .30 | .61 | .52 | .55 | .54 | .47 | .50 | .55 | .55 | - | | |
| 14 Total | .79 | .77 | .58 | .39 | .81 | .79 | .80 | .80 | .69 | .77 | .82 | .84 | .73 | - | |
| Test-Wiseness Test | | | | | | | | | | | | | | | |
| 15 Total | .32 | .30 | .24 | .19 | .32 | .23 | .22 | .24 | .21 | .25 | .28 | .28 | .23 | .31 | - |
| Mean | 49.7 | 49.6 | 49.7 | 49.7 | 49.7 | 49.6 | 49.7 | 49.6 | 49.6 | 49.5 | 49.6 | 49.6 | 49.5 | 49.5 | 49.6 |
| S D | 9.9 | 10.0 | 9.6 | 7.6 | 10.0 | 10.0 | 9.9 | 9.9 | 9.9 | 9.9 | 10.0 | 9.9 | 10.0 | 9.9 | 9.9 |

Note. Scores were standardized across grades 1, 2, and 3 to have a mean of 50 and S D of 10.
For $r \geq .081$, $p < .01$.

Table 9

Intercorrelations of Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests: Total Scores and Subscores
Level II; Grades 4, 5, and 6
(N = 1,728)

| Scores | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCC Test | | | | | | | | | | | | | | | |
| 1 Noun | | | | | | | | | | | | | | | |
| 2 Verb | .87 | | | | | | | | | | | | | | |
| 3 Adjective | .79 | .77 | | | | | | | | | | | | | |
| 4 Adverb | .65 | .64 | .58 | | | | | | | | | | | | |
| 5 Total | .97 | .94 | .87 | .72 | | | | | | | | | | | |
| Wh-Item Test | | | | | | | | | | | | | | | |
| 6 How | .51 | .52 | .46 | .40 | .53 | | | | | | | | | | |
| 7 What(N-P) | .49 | .49 | .44 | .34 | .51 | .45 | | | | | | | | | |
| 8 What(V) | .48 | .49 | .44 | .37 | .51 | .43 | .44 | | | | | | | | |
| 9 When | .48 | .47 | .42 | .35 | .49 | .45 | .41 | .43 | | | | | | | |
| 10 Where | .51 | .48 | .43 | .37 | .51 | .47 | .42 | .41 | .44 | | | | | | |
| 11 Which | .52 | .52 | .46 | .38 | .54 | .49 | .47 | .45 | .47 | .48 | | | | | |
| 12 Who | .52 | .50 | .45 | .38 | .53 | .45 | .43 | .42 | .43 | .44 | .50 | | | | |
| 13 Why | .53 | .52 | .47 | .40 | .55 | .47 | .45 | .44 | .43 | .48 | .47 | .48 | | | |
| 14 Total | .70 | .70 | .62 | .52 | .73 | .73 | .70 | .70 | .70 | .72 | .75 | .71 | .73 | | |
| Test-Wiseness Test | | | | | | | | | | | | | | | |
| 15 Total | .28 | .28 | .27 | .22 | .30 | .22 | .23 | .22 | .23 | .22 | .24 | .25 | .25 | .33 | |
| Mean | 49.7 | 49.7 | 49.6 | 49.6 | 49.6 | 49.6 | 49.5 | 49.5 | 49.5 | 49.5 | 49.5 | 49.6 | 49.5 | 49.6 | 49.5 |
| S D | 9.9 | 9.9 | 9.9 | 9.9 | 9.9 | 10.0 | 9.9 | 9.9 | 10.0 | 10.0 | 9.9 | 10.0 | 10.0 | 10.0 | 9.9 |

Note. Scores were standardized across grades 4, 5, and 6 to have a mean of 50 and S D of 10.
For $r \geq .081$, $p < .01$.

Table 10

Intercorrelations of Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests: Total Scores and Subscores
MCC Level III; Grades 7 and 8
(N = 1,260)[a]

| Scores | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCC Test | | | | | | | | | | | | | | | |
| 1 Noun | - | | | | | | | | | | | | | | |
| 2 Verb | .82 | - | | | | | | | | | | | | | |
| 3 Ajective | .80 | .76 | - | | | | | | | | | | | | |
| 4 Adverb | .62 | .59 | .56 | - | | | | | | | | | | | |
| 5 Total | .96 | .91 | .88 | .69 | - | | | | | | | | | | |
| Wh-Item Test | | | | | | | | | | | | | | | |
| 6 How | .44 | .42 | .36 | .34 | .45 | - | | | | | | | | | |
| 7 What (N-P) | .40 | .39 | .35 | .28 | .42 | .48 | - | | | | | | | | |
| 8 What (V) | .41 | .41 | .36 | .33 | .43 | .46 | .46 | - | | | | | | | |
| 9 When | .43 | .45 | .40 | .32 | .46 | .49 | .48 | .47 | - | | | | | | |
| 10 Where | .46 | .44 | .40 | .33 | .48 | .46 | .44 | .43 | .45 | - | | | | | |
| 11 Which | .43 | .42 | .38 | .32 | .45 | .46 | .48 | .45 | .49 | .45 | - | | | | |
| 12 Who | .47 | .48 | .41 | .35 | .50 | .52 | .48 | .49 | .53 | .53 | .51 | - | | | |
| 13 Why | .49 | .44 | .42 | .34 | .49 | .47 | .48 | .50 | .50 | .48 | .48 | .51 | - | | |
| 14 Total | .60 | .59 | .52 | .44 | .62 | .73 | .73 | .71 | .74 | .72 | .73 | .77 | .75 | - | |
| Test-Wiseness Test | | | | | | | | | | | | | | | |
| 15 Total | .24 | .19 | .23 | .16 | .24 | .11 | .13 | .11 | .17 | .16 | .12 | .13 | .12 | .18 | - |
| Mean | 49.66 | 49.56 | 49.58 | 49.71 | 49.64 | 49.54 | 49.67 | 49.62 | 49.56 | 49.64 | 49.59 | 49.55 | 49.59 | 49.65 | 49.59 |
| S D | 9.87 | 9.87 | 9.86 | 9.90 | 9.85 | 9.91 | 9.84 | 9.89 | 9.90 | 9.89 | 9.88 | 9.78 | 9.93 | 9.84 | 9.95 |

Note: Scores were standardized across grades 7 and 8, to have a mean of 50 and S D of 10.
For $r \geq .081$, $p < .01$.
[a] Grade 9 is not included.

organized by test level on the wh-item and cloze tests (Level I for grades 1-3; Level II for grades 4-6; and Level III for grades 7-9).[4]

Test Comparability

The issue of test comparability has two important facets of concern here. The first is the purely technical one of determining whether the various test forms constructed for the research were reasonably parallel and thus could be combined in the analyses by subsample level.

The other facet of test comparability has to do with determining the extent to which the wh-item and multiple-choice cloze test measure the same thing. These tests were originally constructed as alternate methods of measuring the same construct--literal comprehension. Tables 8-10 are relevant to the latter concern, though these data are not a basis for strong inferences about the parallelism of the constructs measured by the wh-item and multiple-choice cloze tests. The intercorrelations in Tables 8-10 are presented in three sections. For example, Table 8 presents the intercorrelations of the deletion type subscores for the Multiple-Choice Cloze Exercises in the upper left-hand traingular section of the matrix; the wh-item subscores for the Wh-item Test are in the lower-right triangle; and the intercorrelations of the subscores of the two tests are presented in the lower-left rectangular section of the table. The correlations of the subscores of both tests with the total test scores are given in rows 5 and 14 of the matrix. The correlations of all scores with the test-wiseness score are given in row 15 of the table. These data support the following observations in the Level I sample:

---

[4] Table 10 is for test Level III but the data are for grade 8 only. Not all of the correlational analyses were completed in the 7-9 sample at the time of this writing.

1. The deletion type subscores for the multiple-choice cloze test form are moderately to highly intercorrelated, except for the adverb score which tends to have low correlations with the other subscores.

2. The noun score correlates highest with all other deletion type subscores and with the total test. (The correlation of nouns with the total test is particularly high, but this is due to the fact that the test is largely composed of noun responses at Level I).

3. Although the wh-subscores are each based on only a few items (the average equals 3.5), these subscores are consistently and moderately interrelated.

4. The wh-item subscores generally correlate at about the same level with the total wh-test score.

5. The correlations of noun or verb scores with the wh-item subscores are generally as high as the intercorrelations of the wh-item subscores.

6. The correlation between the total scores on both the multiple-choice cloze and wh-item tests is .81, a level at which either test is fairly predictable from the other. (If these tests were two norm-referenced measures of reading comprehension, this correlation would be interpreted as a very acceptable validity coefficient.)

7. The test-wiseness score has a significant but very low correlation with the other test scores.

The foregoing observations hold for the other levels of the multiple-choice cloze and wh-item tests given in Tables 9 and 10, with a few important qualifications. The correlations of adjective and adverb scores with subtest scores and total test scores on the multiple-choice cloze test improve across grades 4-8, due apparently to changes in the composition of passages at these more advanced reading levels. (The mean and variance of advert and adjective subscores increase, but the total test is still predominantly noun and verb.) The correlations between subscores and total scores on the wh-item and multiple-choice cloze tests tend to decrease across grade levels, with this decrease being most marked in the grade 7-8 sample ($\underline{r}$ between total scores is .62 as compared to .81 in the Level I subsample).

These data lead to the conclusion that the two types of literal comprehension tests assembled for the research are generally consistent within themselves across the study sample. The pattern of the intercorrelations within and between test types is further consistent with the interpretation that the subscores in either test contribute to a factor that is common to both tests. However, the commonality that appears evident between the two tests is substantially reduced in the upper grade samples, an effect that is apparently attributable to the shapes of the score distributions at these levels and reduction in range of talent.

The other factor of test comparability of interest here may be evaluated in part by reference to Table 11, which presents the means and standard deviations for each test form in the wh-item and multiple-choice cloze formats. In addition to these data, the proportions of correct responses for each item in each test form were arrayed and visually inspected as were the raw score distributions on each test form. Inspections of the means and standard deviations of each set of test forms shows that the difference between any pair of test-form means is relatively small (e.g., for the Level I multiple-choice cloze test forms, the largest difference between means is less than one-third of the average standard deviation; for the Level I wh-item test, the average difference between means is less than one-fourth of the average standard deviation). These data, taken together with the examination of the score distributions on each wh-item and multiple-choice cloze test form, indicate that the various forms of each test type in a test level were fairly comparable. Given the other properties of each test form (e.g., progressive ordering of passages by readability level), it then appears that the various test forms of each type in a level would result in fairly comparable scaling of students at each level of the wh-item and multiple-choice cloze tests.

Table 11

Means and Standard Deviations for the Multiple-Choice Cloze and Wh-Item Tests

| Level | Multiple Choice Cloze Form | N | X | S.D. | Wh-Item Test Form | N | X | S.D. |
|---|---|---|---|---|---|---|---|---|
| I (Grades 1, 2, 3) | 1 | 128 | 21.03 | 10.55 | 37 | 127 | 18.80 | 7.15 |
| | 2 | 126 | 20.26 | 10.49 | 38 | 124 | 19.52 | 7.32 |
| | 3 | 130 | 21.51 | 10.59 | 39 | 126 | 19.57 | 7.29 |
| | 4 | 124 | 22.44 | 11.34 | 40 | 121 | 19.02 | 7.30 |
| | 5 | 126 | 23.06 | 11.24 | 41 | 119 | 18 43 | 7.72 |
| | 6 | 126 | 19.71 | 9.24 | 42 | 122 | 19.17 | 7.00 |
| | 7 | 127 | 21.47 | 11.22 | 43 | 124 | 19.05 | 7.28 |
| | 8 | 127 | 18.84 | 10.40 | 44 | 124 | 19.20 | 7.53 |
| | 9 | 129 | 21.98 | 11.31 | 45 | 131 | 19.10 | 7.69 |
| | 10 | 127 | 20.47 | 10.43 | 46 | 123 | 19.19 | 7.47 |
| | 11 | 120 | 23.39 | 11.25 | 47 | 121 | 18.65 | 7.18 |
| | 12 | 123 | 22.67 | 11.41 | 48 | 121 | 20.12 | 7.69 |
| II (Grades 4, 5, 6) | 13 | 147 | 41.46 | 11.45 | 49 | 147 | 22.74 | 5.57 |
| | 14 | 151 | 40.01 | 14.11 | 50 | 153 | 21.95 | 5.46 |
| | 15 | 153 | 38.73 | 12.51 | 51 | 148 | 22.72 | 4.85 |
| | 16 | 152 | 40.99 | 11.62 | 52 | 152 | 22.74 | 5.66 |
| | 17 | 146 | 42.18 | 12.60 | 53 | 145 | 23.52 | 5.46 |
| | 18 | 151 | 36.35 | 11.03 | 54 | 144 | 23.19 | 5.45 |
| | 19 | 152 | 41.80 | 13.48 | 55 | 145 | 22.96 | 5.00 |
| | 20 | 148 | 42.00 | 12.08 | 56 | 149 | 22.60 | 5.96 |
| | 21 | 152 | 41.39 | 11.37 | 57 | 147 | 20.76 | 6.11 |
| | 22 | 152 | 39.63 | 13.57 | 58 | 148 | 22.19 | 5.86 |
| | 23 | 148 | 41.72 | 12.99 | 59 | 157 | 23.76 | 5.51 |
| | 24 | 149 | 39.01 | 13.32 | 60 | 145 | 21.87 | 5.73 |
| III (Grades 7, 8, 9) | 25 | 167 | 36.60 | 12.53 | 61 | 163 | 23.81 | 5.54 |
| | 26 | 164 | 36.44 | 11.69 | 62 | 162 | 23.89 | 7.01 |
| | 27 | 160 | 38.86 | 14.33 | 63 | 164 | 24.25 | 5.79 |
| | 28 | 161 | 40.47 | 12.82 | 64 | 161 | 23.89 | 4.83 |
| | 29 | 158 | 39.17 | 11.52 | 65 | 165 | 23.53 | 4.75 |
| | 30 | 165 | 42.54 | 13.35 | 66 | 166 | 21.20 | 6.20 |
| | 31 | 158 | 39.46 | 12.45 | 67 | 154 | 24.88 | 4.85 |
| | 32 | 163 | 37.07 | 12.01 | 68 | 163 | 22.40 | 5.65 |
| | 33 | 166 | 37.38 | 11.98 | 69 | 164 | 24.02 | 4.99 |
| | 34 | 159 | 38.08 | 13.60 | 70 | 156 | 22.01 | 5.31 |
| | 35 | 163 | 37.82 | 13.18 | 71 | 163 | 23.16 | 5.94 |
| | 36 | 165 | 41.82 | 12.56 | 72 | 154 | 22.03 | 6.90 |

Test Reliability

The reliability of a test is of interest because it estimates the amount of random error contained in a test score. The validity coefficients reported here for the wh-item, multiple-choice cloze, and test-wiseness tests cannot be interpreted satisfactorily without an appropriate estimate of measurement error. The reliability statistic selected for this purpose was the Kuder-Richardson Formula 20 or KR-20 (Kuder and Richardson, 1937), which is a special case of the Hoyt (1941) or Cronbach (1951) coefficients of equivalence when test items are scored dichotomously. As used here, KR-20 reflects directly on several other properties of the cloze and wh-item tests. The formula provides an estimate of the homogeneity of the items in the test or the proportion of test variance attributable to the first general factor in the test. In addition, the KR-20 formula provides a good estimate of the short-term stability of the test. And, since the KR-20 is available here on a large number of test forms systematically assembled from the cloze and wh-item passage pools, the average or median value of the KR-20 is also a basis for judging the reliability of the process of test generation.

Accordingly, the KR-20 also reflects on the validity of the wh-item and multiple-choice cloze tests. The description of the construct indicates the test is a measure of a homogeneous trait that should be stable over short periods of time. Inspection of the KR-20 coefficients in Table 12 for the multiple-choice cloze and wh-item tests indicates that reliability expectations for these tests were confirmed at a very high level. The median KR-20 value for the 12 cloze test forms at each test level was .96; for the wh-item test, the KR-20 ranged from .91 to .93. These data provide further support for the conclusion that each of the two measures of

Table 12

Kuder-Richardson Formula 20 Reliability Coefficients for the Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests

| Level | Multiple-Choice Cloze Test | | | | | Wh-Item Test | | | | | Test-Wiseness Measure | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Form | N | I | KR-20 | SE | Form | N | I | KR-20 | SE | Form | N | I | KR-20 |
| I (Grades 1,2,3) | 1 | 128 | 41 | .94 | 1.73 | 37 | 127 | 30 | .92 | 2.02 | 37 | 127 | 12 | .76 |
| | 2 | 126 | 41 | .95 | 1.64 | 38 | 124 | 30 | .94 | 1.79 | 38 | 124 | 12 | .68 |
| | 3 | 130 | 41 | .96 | 1.46 | 39 | 126 | 30 | .90 | 2.30 | 39 | 126 | 12 | .68 |
| | 4 | 124 | 41 | .96 | 1.46 | 40 | 121 | 30 | .90 | 2.31 | 40 | 121 | 12 | .47 |
| | 5 | 126 | 41 | .95 | 1.73 | 41 | 119 | 30 | .91 | 2.32 | 41 | 119 | 12 | .67 |
| | 6 | 126 | 39 | .95 | 1.57 | 42 | 122 | 30 | .91 | 2.10 | 42 | 122 | 12 | .79 |
| | 7 | 127 | 41 | .96 | 1.45 | 43 | 124 | 30 | .93 | 1.92 | 43 | 124 | 12 | .68 |
| | 8 | 127 | 39 | .96 | 1.51 | 44 | 124 | 30 | .90 | 2.38 | 44 | 124 | 12 | .51 |
| | 9 | 129 | 41 | .97 | 1.33 | 45 | 131 | 30 | .90 | 2.43 | 45 | 131 | 12 | .26 |
| | 10 | 127 | 41 | .96 | 1.49 | 46 | 123 | 30 | .91 | 2.24 | 46 | 123 | 12 | .78 |
| | 11 | 120 | 41 | .96 | 1.43 | 47 | 121 | 30 | .94 | 1.76 | 47 | 121 | 12 | .13 |
| | 12 | 123 | 41 | .96 | 1.54 | 48 | 121 | 30 | .92 | 2.18 | 48 | 121 | 12 | .76 |
| | Median | | | .96 | 1.49 | | | | .91 | 2.21 | | | | .68 |
| II (Grades 4,5,6) | 13 | 147 | 60 | .97 | 1.98 | 49 | 147 | 30 | .93 | 1.47 | 49 | 147 | 12 | .50 |
| | 14 | 152 | 60 | .96 | 2.82 | 50 | 153 | 30 | .93 | 1.44 | 50 | 152 | 12 | .75 |
| | 15 | 153 | 60 | .96 | 2.50 | 51 | 148 | 30 | .90 | 1.53 | 51 | 148 | 12 | .68 |
| | 16 | 152 | 60 | .96 | 2.32 | 52 | 152 | 30 | .86 | 2.11 | 52 | 152 | 12 | .74 |
| | 17 | 146 | 60 | .97 | 2.18 | 53 | 145 | 30 | .93 | 1.44 | 53 | 145 | 12 | .29 |
| | 18 | 151 | 60 | .94 | 2.69 | 54 | 144 | 30 | .92 | 1.54 | 54 | 144 | 12 | .76 |
| | 19 | 152 | 60 | .97 | 2.33 | 55 | 145 | 30 | .85 | 1.94 | 55 | 145 | 12 | .46 |
| | 20 | 148 | 60 | .95 | 2.69 | 56 | 149 | 30 | .95 | 1.33 | 56 | 149 | 12 | .73 |
| | 21 | 152 | 60 | .95 | 2.53 | 57 | 147 | 30 | .94 | 1.50 | 57 | 147 | 12 | .71 |
| | 22 | 152 | 60 | .97 | 2.35 | 58 | 148 | 30 | .91 | 1.76 | 58 | 148 | 12 | .58 |
| | 23 | 148 | 60 | .97 | 2.25 | 59 | 157 | 30 | .94 | 1.35 | 59 | 157 | 12 | .66 |
| | 24 | 149 | 60 | .95 | 2.97 | 60 | 147 | 30 | .93 | 1.51 | 60 | 145 | 12 | .76 |
| | Median | | | .96 | 2.35 | | | | .93 | | | | | .70 |

Table 12 (Continued)

| Level | Multiple-Choice Cloze Test | | | | | Wh-Item Test | | | | | Test-Wiseness Measure | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Form | N | I | KR-20 | SE | Form | N | I | KR-20 | SE | Form | N | I | KR-20 |
| III(Grades 7,8,9) | 25 | 167 | 60 | .96 | 2.51 | 61 | 163 | 30 | .91 | 1.66 | 61 | 163 | 12 | .55 |
| | 26 | 164 | 60 | .95 | 2.61 | 62 | 162 | 30 | .94 | 1.71 | 62 | 162 | 12 | .70 |
| | 27 | 160 | 60 | .96 | 2.87 | 63 | 164 | 30 | .96 | 1.16 | 63 | 164 | 12 | .77 |
| | 28 | 161 | 60 | .97 | 2.22 | 64 | 161 | 30 | .89 | 1.60 | 64 | 161 | 12 | .78 |
| | 29 | 158 | 60 | .96 | 2.30 | 65 | 165 | 30 | .89 | 1.57 | 65 | 165 | 12 | .59 |
| | 30 | 165 | 60 | .97 | 2.31 | 66 | 166 | 30 | .94 | 1.52 | 66 | 166 | 12 | .77 |
| | 31 | 158 | 60 | .95 | 2.78 | 67 | 154 | 30 | .96 | 0.97 | 67 | 154 | 12 | .75 |
| | 32 | 163 | 60 | .96 | 2.40 | 68 | 163 | 30 | .90 | 1.78 | 68 | 163 | 12 | .79 |
| | 33 | 166 | 60 | .95 | 2.67 | 69 | 164 | 30 | .96 | 0.99 | 69 | 164 | 12 | .77 |
| | 34 | 159 | 60 | .95 | 3.03 | 70 | 156 | 30 | .93 | 1.40 | 70 | 156 | 12 | .70 |
| | 35 | 163 | 60 | .97 | 2.28 | 71 | 163 | 30 | .95 | 1.32 | 71 | 163 | 12 | .75 |
| | 36 | 165 | 60 | .97 | 2.17 | 72 | 154 | 30 | .95 | 1.56 | 72 | 154 | 12 | .72 |
| | Median | | | .96 | 2.40 | | | | .94 | 1.54 | | | | .75 |
| Overall | Median | | | .96 | | | | | .92 | | | | | .72 |
| | Mean | | | .96 | | | | | .92 | | | | | .65 |
| | Range | | | .94-.97 | | | | | .85-.96 | | | | | .13-.79 |

Note. N = number of subjects.
I = number of items.

literal comprehension is a highly reliable estimate of a single homogeneous trait. The tests assembled to measure the trait should thus scale individuals similarly over short periods of time, and it appears that similar tests for measuring this trait can be repeatedly assembled.[5]

A final point of comment concerns the reliability of the test-wiseness test forms assembled for each test level. Across all test levels, 4 of these test forms had low reliabilities (Forms 40, 45, 47, and 53), but given the shortness of these test forms and their unusual composition, the KR-20's are surprisingly high. These results are similar to Tuinman's (1973) for a similar test he referred to as a measure of passage independence.

---

[5] This inference is supported directly and indirectly by the data reported here and by the consistency of the results reported in the next section of this report.

## Construct Validation

The principal focus of construct validation is the explication of the network of interrelated concepts that define the trait in question and the conditions and interpretations that surround its measurement (Messick, 1975). One accepted approach to construct validation is to examine the convergence between the principal measure in question and other measures that are indicators of the same construct or can be discriminated from the construct. Another approach is to examine the convergence of two alternate measures of the same construct, obtained by dissimilar methods, with other measures that are indicators of the same construct or can be discriminated from the construct (Cronbach, 1971).

In the present analysis, the first approach consisted of analyzing expected levels of correlation between multiple-choice cloze scores and various other test scores available in the study--the wh-item test, the CAT skill scores and subtest scores, language and non-language IQ test scores, and test-wiseness scores. In the second approach, the convergence of the multiple-choice cloze and wh-item tests as measures of literal comprehension was evaluated in a simultaneous comparison of the values of the correlations of both of these tests with the other test criteria available in the study.

In the first approach, convergent validity is evaluated by the general consistency with which predictions are confirmed in terms of relative levels of correlation between the multiple-choice cloze test and other measures in the study considered to be similar indicators of the construct of literal comprehension or related or unrelated indicators of other constructs. Some measures are expected to correlate relatively highly with the multiple-choice cloze test, others relatively moderately, and still others are

predicted to have a low correlation. In the second approach, convergent validity is determined by reference to the size of the differences in the absolute values of the correlations of the multiple-choice cloze and wh-item tests with the other test criteria available in the study. These absolute differences are expected to be small, consistent from measure to measure, and generalizable across levels of the study population.

In general, the predictions indicate that the multiple-choice cloze test is relatively highly related to the wh-item test, unrelated to the measure of test-wiseness, moderately related to the measure of language IQ, and related in varying degrees to the CAT, depending on the "apparent meaning" of a skill or subtest score denoted by the test label. Both the labels and specific item content of these CAT skill and subtest scores present some difficulties in interpretation, inevitably leading to ambiguity concerning whether a given test is a similar measure of the construct or has some other relationship with literal comprehension. Where these ambiguities arise in the present analysis, disconfirming predictions, an attempt is made to resolve the problem by examination of the test items in question. Such an analysis, however, is potentially fraught with the usual problems of all post hoc analyses. That is, some explanation of the disconfirming event can usually be found, and for this reason, a post hoc analysis must be taken as exploratory or hypothetical.

Convergent Validity of the Multiple-Choice Cloze

It was assumed that predictions about the correlations between the multiple-choice cloze test and the other test criteria available in the study could be made on the basis of the construct definition of literal comprehension and the information at hand defining the content of the test criteria. In order to understand the rationale for these predictions, it is instructive to review briefly the construct definition and the availability of information that constitutes an adequate definition of each of the

various potential indicators or non-indicators of it.

The construct: the multiple-choice cloze test. The construct, literal comprehension, and its relationship with the multiple-choice cloze test are stated in detail elsewhere (O'Reilly et al., 1976). In brief, literal comprehension is the apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse" (Katz and Fodor, 1967, p. 172). The multiple-choice cloze test accesses these grammatical and semantic relations by systematically deleting nouns, verbs, and modifiers[6] from a segment of written discourse, and then placing the deleted words in sets of responses where the distractors are all grammatically plausible but semantically implausible. It is hypothesized that students will have no difficulty in selecting the only word which is grammatically and semantically plausible if they can apprehend "the grammatical and semantic relations which obtain within and among the sentences of the discourse." The distractors, in other words, do not function as traditional distractors--do not, in fact, "distract"--until the syntactic and semantic complexity of the discourse exceeds the students' psycholinguistic competence. The test is designed, therefore, to discriminate between a specifiable set of interactions--called literal comprehension--between student and text, and another specifiable set of interactions between student and text called no comprehension. The test is designed, that is, to measure literal comprehension or no comprehension and nothing else. The interactions between student and test--the extensiveness of the processing of the grammatical and semantic relations in the text--are carefully controlled by the type and rate of deletion and the distractor selection procedure. The item type is hypothesized to access only literal meaning; it should access no nuances

[6] Only nouns and verbs are deleted in grade 1 and 2 materials.

of meaning and no other semantic interrelationships than those clearly signaled in the grammatical and semantic relations of the text.

The construct: the wh-item test. The wh-item test is designed to access the same grammatical and semantic relations of a given text. Like the multiple-choice cloze, the wh-item test is considered an indicator of the construct, literal comprehension. The wh-item accesses the grammatical and semantic relations of a text by deleting immediate constituents in clauses of the sentences in the text, replacing them with the appropriate wh-words (who, what, which, where, when, how, or why), and then transforming the clauses into questions.

The wh-item, then, is the traditional question type teachers use to direct student attention to salient features of the text, and correct answers to such questions are usually considered evidence of literal comprehension. The primary difference between such traditional questions and the wh-item test is the systematic way in which the wh-item is written. Such systematization makes it possible to specify and control, to a greater degree than possible with traditional test questions, the interactions between the features of the text and the psycholinguistic competence of the student.

Since the wh-item test has some claim to a specifiable relationship with the construct definition, it is considered the preferred or least ambiguous criterion measure for the multiple-choice cloze in the analysis that follows. The primary difficulty with using the wh-item as a criterion measure is that it is subject to a form of test-wiseness. In brief, it is possible for a student with minimal syntactic competence to locate the correct answers to wh-items in the text without understanding

what the question or the text means. However, as will be seen throughout the course of the analyses, there is no reason to suspect that the wh-item test was actually subject to this form of test-wiseness in the study samples.

The construct: other criterion measures. The test-wiseness measure used in this analysis was briefly described at the outset of this paper and is not to be confused with the form of test-wiseness just discussed. The test-wiseness measure used in the analysis is a preliminary effort to determine the passage dependence of the wh-items. Passage dependency is also crucial to the construct, literal comprehension. That is, the test items must access only the grammatical and semantic relations which obtain within and among the sentences of the discourse.

The language and non-language IQ scores reported in the following analysis come from the Short Form Test of Academic Aptitude. In the development of the construct definition, it was hypothesized that language IQ scores should only correlate moderately with scores from tests of literal comprehension, since the literal level of comprehension requires little of the inferential and related reasoning processes so characteristic of measures of verbal intelligence. Non-language IQ scores would seem to have little or no relation to the construct of literal comprehension and should thus correlate to a lesser degree with the multiple-choice cloze than language IQ.

The problems that exist in specifying the relationship between the construct definition of literal comprehension and each of the skill and subscores of the CAT have already been noted. Given the lack of an explicit statement defining the psycholinguistic meaning of each CAT skill and subscore used in the analysis, it was necessary to define the relationship of these

skills with the construct by recourse to the test score labels and the meager skill or subscore descriptions given in the CAT Test Coordinator's Handbook (1970).

Cat skill scores: predicted correlational levels. The predicted correlational levels between multiple-choice cloze scores on the one hand, and the wh-item test, CAT skills, and language and non-language IQ scores on the other, are presented in Table 13. These predictions are based upon the relative degree to which it is expected the different measures will converge upon the construct, literal comprehension. The predictions for the CAT scores have been based on the labels or brief descriptions attached to a skill score or subscore, and the consistency of the application of the labels is assumed.

Table 13

Expected Levels of Correlation of Multiple-Choice Cloze Scores with Wh-Item Test Scores, California Achievement Test Skill Scores, Test-Wiseness Scores, and Language and Non-Language IQ Scores

| Lowest (.00-.29) | Medial (.30-.54) | Highest (.55+) |
|---|---|---|
| Test-Wiseness | Language Usage | Wh-Item Test Scores |
| Language Mechanics | Language IQ | Reading Vocabulary |
| Non-Language IQ | | Reading Comprehension |

The highest predicted levels of correlation as evidenced by Table 13, are between the multiple-choice cloze scores and the wh-item test scores, the CAT reading vocabulary, and the CAT comprehension scores. The crucial correlation is, of course, between the multiple-choice cloze and the wh-item test, the preferred criterion measure. The two experimental tests must correlate highly with each other. A strong prediction is also made about the correlations between the multiple-choice cloze scores and scores from

the test-wiseness measure, for reasons already cited. The remaining predictions are less strong because the relationship between the construct and the remaining measures is not so clear. Language mechanics and non-verbal IQ scores are expected to have low correlations with multiple-choice cloze scores while language usage and language IQ scores should fall in the medial range.

Also, implicit in Table 13 is the assumption that the multiple-choice cloze test will behave consistently across grade levels as long as passages are properly matched in readability with the psycholinguistic competence of the students. Such consistent behavior is crucial to the possibility of using the same item type to measure literal comprehension, regardless of the content of the test passages or the reading ability of the student. Hence, the predictions in Table 13 are not made by test level.

Actual correlational levels. The actual correlational levels between multiple-choice cloze scores and the criterion measures previously discussed are given in Table 14. When the predicted correlational level matches the actual correlational level, the actual correlation is underlined. As can be seen in Table 14, 17 out of 30 predictions were confirmed. Several others, were close to predicted levels. More importantly, beyond the consistent pattern of confirmation, the crucial correlational levels--namely the wh-item test, the CAT vocabulary, the CAT comprehension, and the test-wiseness scores--were generally confirmed. However, there are disconfirmations at Level IV, along with a generally tendency for the values of the correlations to fall off. Preceding sections of this chapter have already analyzed this phenomenon. What is important here is to note that the correlational pattern in Level IV remains consistent with preceding CAT levels in spite of the reduced values. No attempt will be made to

explain the failed predictions in the case of language usage, language mechanics, language and non-language IQ until the actual test items are examined and ordered in relation to the construct, literal comprehension. Further analysis of correlational levels between CAT skill scores and multiple-choice cloze scores is taken up in finer detail in the next section by breaking the CAT skill scores into their component subscores.

Table 14

Actual Zero-Order Correlations of Multiple-Choice Cloze Scores with Wh-Item Test Scores, California Achievement Test Skill Scores, Language and Non-Language IQ Scores, and Test-Wiseness Scores

| | CAT Test Level | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| Wh-Item Test | .64 | .73[a] | .73 | .62 |
| Vocabulary | .64 | .76 | .69 | .50 |
| Comprehension | .61 | .79 | .71 | .47 |
| Language Usage | .50 | .73 | .57 | .41 |
| Language Mechanics | .55 | .71 | .68 | .50[b] |
| Language IQ | .34 | .55 | .62 | - |
| Non-Language IQ | .43 | .49 | .52 | - |
| Test-Wiseness | .23 | .26 | .29 | .24 |

Note. Underlined values are within the predicted correlational levels.

[a]The r of wh- vs. cloze in the combined grades 1 to 3 is .81.

[b]The Level IV sample did not receive the IQ test.

CAT subscores: predicted correlational levels. Table 15 presents the predicted correlational levels between multiple-choice cloze scores and subscores on the CAT. Again, the predictions are expected to hold regardless of grade levels or variations in content of test passages. Predictions are also based on the relationships between the construct, literal com-

prehension, and the CAT subscores, insofar as that can be determined from the meager descriptions of the subskills in the CAT Test Coordinator's Handbook (1970). For purposes of this analysis, it is again assumed that the CAT labels are applied consistently to test items.

Table 15

Predicted Levels of Correlations between Multiple-Choice Cloze Test Scores and California Achievement Test Subscores

| | Lowest (.00-.29) | Medial (.30-.54) | Highest (.55+) |
|---|---|---|---|
| Vocabulary subscores | Sentence-Picture Assoc.<br>Beginning Sounds<br>Ending Sounds<br>Letter Recognition<br>Word Form | Picture-Word Assoc.<br>Word Recognition | Words in Context |
| Comprehension subscores | Inferences | Relationships<br>Generalizations<br>Reading-General<br>Reading-Soc. Studies<br>Reading-Science<br>Reading-Math. | Facts<br>Interpretations |
| Language subscores | Sentence Structures<br>Sentence Parts and Functions<br>Transformations | Standard English | |

Table 15 indicates that the most important correlations are between the multiple-choice cloze scores and the CAT Words in Context, Facts, and Interpretation subscores. They are expected to correlate highest with the multiple-choice cloze since they seem to access "the grammatical and semantic relations which obtain within and among the sentences of the discourse." Strong predictions are also made that the CAT Inferences subscores will correlate least with multiple-choice cloze scores since inferential and related reasoning processes require reasoning beyond the literal meaning

of the test passages. Such items are also liable to be least passage dependent, contrary to the demands of the construct, literal comprehension. Another strong prediction is that those CAT items that measure phonological skills (Sentence-Picture Association, Beginning Sounds, and Ending Sounds) will correlate lowest with multiple-choice cloze scores since the model of reading as a constructive language process given in the full report of this research (O'Reilly et al., 1976) posits that phonological processes are ordinarily bypassed in processing written discourse. The remaining subskills are parceled out according to their apparent relationship with the construct. The reading scores in the subject areas on the CAT are expected to correlate in the medial range with multiple-choice cloze scores since the subject area scores seem to subsume the full range of comprehension subskills from facts to inferences.

Actual correlational levels. Table 16 gives the actual correlational levels between multiple-choice cloze scores and CAT subscores. As Table 16 indicates, 28 of 47 correlations fall within the predicted levels. An additional five subscores--including Interpetation, Levels I and IV--are very close to predicted levels. The consistent pattern, then, is to confirm predictions about correlational levels with CAT Words in Context, Facts, and Interpretation subscores and Test-Wiseness scores. As noted previously, correlational values generally fall off in Level IV as a result of a reduction in the range of student ability represented at that level; consequently, there are disconfirmations of predicted correlational levels for Facts and Interpretation at Level IV. The general pattern, however, is still evident in Level IV. That is, the highest correlations at Level IV are Words in Context and Interpretation.

Table 16

Actual Zero-Order Correlations between Multiple Choice Cloze Test Score and California Achievement Test Subscores by CAT Level

| | CAT Level | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| **Vocabulary Subscores** | | | | |
| Sentence-Picture Association | .06 | | | |
| Beginning Sounds | .43 | | | |
| Ending Sounds | .35 | | | |
| Letter Recognition | .11 | | | |
| Word Form | .27 | | | |
| Picture-Word Association | .53 | | | |
| Word Recognition | .37 | .56 | | |
| Words In Context | .65 | .76 | .70 | .57 |
| **Comprehension Subscores** | | | | |
| Facts | .60 | .76 | .67 | .45[a] |
| Interpretation | .52 | .73 | .70[a] | .50 |
| Relationships | | | .43 | .38 |
| Generalizations | | .73 | .51[a] | .42 |
| Inferences | .48 | .73 | .35 | .44 |
| Reading-General | | | .62 | .46 |
| Reading-Social Studies | | | .58 | .46 |
| Reading-Science | | | .58 | .42 |
| Reading-Mathematics | | | .51 | .40 |
| **Language Subscores** | | | | |
| Standard English | .44 | .73 | .46[a] | .31 |
| Sentence Structure | | | .37 | .42 |
| Sentence Parts and Functions | | | .31[a] | .45 |
| Transformation | | | .28 | .22 |

Note. Underlined values were within level of correlation predicted.

[a]Five or fewer items on the CAT

The results, then, tend to substantiate assumptions about the relationships between the construct, literal comprehension, its principal indicator, the multiple-choice cloze, and CAT subscores.

Inconsistent correlational levels. Despite the general tendency to corroborate the assumptions behind the predictions, there are a considerable number of disconfirmations evident in Table 16 that require further analysis, even in this preliminary investigation. The most notable inconsistency is the unexpectedly high level of correlation between multiple-choice cloze scores and CAT Inferences subscores, especially at Levels I and II. In the following discussion of these inconsistencies, the assumptions behind the predictions are examined in more detail, and then the CAT inference items themselves are reviewed in relation to the construct, literal comprehension.

Any interpretation of written discourse involves "inferential" processes. According to the model of reading guiding this research, meaning is not in the text; rather, meaning is in the reader and the writer, and what appears on the printed page is only an approximation of the meaning intended by the writer or apprehended by the reader. The text contains only orthographic clues to meaning: The reader must "infer" grammatical and semantic relations in and among the sentences of the discourse from the linguistic clues to such relations in the text. But these "inferential" processes are language-specific; that is, they are part of the grammar of the language and are, therefore, well within the processes of the construct, literal comprehension--the apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse."

On the other hand, "inferences," as commonly understood in

educational psychology, refer to deductive and related formal reasoning processes that are not part of the grammar of a language. Such inferential processes are quite beyond the psycholinguistic processes of literal comprehension, and require not only the apprehension of the literal meaning of a text but also the apprehension of the relationship between literal meaning and other information not in evidence in the text. As noted previously such inferential processes tend to subordinate the information in the text to extra-textual information, thus reducing the loading of literal comprehension in the test and the passage dependency of the test items. It was with these kinds of inferential processes in mind that the predictions regarding correlational levels between multiple-choice cloze scores and CAT Inferences subscores were made. The predicted correlational levels were low, but, as noted above, the actual levels were medial and high. The CAT items were examined in an attempt to explain the inconsistency.

An examination of the CAT test items at Levels I and II implies a vague, global notion of inference. That is, test items that vary greatly in the kinds of demands they make upon the reasoning processes of the student are all subsumed under the label, "Inferences." Item numbers 3, 16, 20, 21, and 24 at CAT Level I, for instance, are all labeled "Inferences" but make very low level demands on student reasoning processes. Item 3 is characteristic:

> A small boy named Henry lived in the city. He had a pet dog, a kitten, and two birds in his home. Henry liked to play with the dog best.
>
> 3. How many pets did Henry have?
>
> o one
> o two
> o three
> o four

The student is asked to demonstrate some ability to coordinate arithmetical skills with literal comprehension reading skills. Item 5, however, makes quite different demands upon the student's inferential abilities:

5. When Henry takes care of his animals, he is

- o busy
- o lazy
- o sorry
- o worried

In the first place, the item stem introduces subordination, thus demanding more linguistic prowess than the sentences in the text. Secondly, the information necessary to make a judgement among the responses is not clearly stated in the text. Thirdly, the semantic complexity of the responses exceeds the level of vocabulary in the text and in the other sets of responses accompanying the test passages. Fourthly, the item is obviously passage independent. Given "animals" (plural) in the item stem, "busy" would merely be descriptive of someone taking care of them. But any of the other distractors is plausible. There is no information in the text that makes "busy" any more correct than the other responses. The correct response can be "inferred" just as easily from the item stem as from the test passage.

In summary, then, there is a wide range of inferential skills subsumed under the label, "Inferences" in CAT Level I. Items 3, 16, 20, 21, and 24 make minimal demands on the reasoning powers of the student, emphasizing instead the grammatical and semantic relations within the discourse. Items 5, 10, and 11, on the other hand, demand that the student reason beyond the grammatical and semantic relations within the discourse. The majority of the items, therefore, emphasize literal comprehension in spite of the label, "Inferences" and hence correlate higher with the multiple-choice cloze than predicted.

The same pattern is repeated at CAT Level II. There are eight items labeled "Inferences." Item 16 is similar to item 3 on CAT Level I previously discussed, and requires only an understanding of the concept, "more than one," and an ability to relate the concept to the passage:

> The children in Mrs. Kim's room were talking about how to make scrapbooks. Eva said, "I will bring some pictures." "I will bring some scissors," Monty said. Marie said, "and I will bring some paper."
>
> The children decided they would need more paste than they had. To make paste they would need water, flour, and salt. Eva said, "I will bring a pan to mix them in."
>
> 16. Who will bring more than one thing?
>
> o Eva
> o Marie
> o Monty
> o Mrs. Kim

Item 15, on the other hand, like item 5 in CAT Level I, requires a characterization of the action of the paragraph beyond its grammatical and semantic relations:

> 15. The children were
>
> o busy
> o lazy
> o playing
> o tired

Item 30 is passage independent:

> 30. The windmill was turned by
>
> o a motor
> o a pump
> o the water
> o the wind

If the student has the semantic knowledge, he can obviously answer such a question without reading the text. Even without such knowledge, a student can "infer" the answer from the item stem. Item 40 is also passage independent:

40. A polar bear's hairy feet are especially useful on

- o ice and snow
- o rocky ground
- o sandy beaches
- o sharp stones

These passage independent items, however, access such a limited semantic knowledge that a student who understands the vocabulary of the test item will have no difficulty choosing the correct answer. In other words, such test items, even though they are labeled "Inferences" seem to access even lower level psycholinguistic processes than literal comprehension. Items 30 and 40 are little more than simple vocabulary tests; "windmill" is associated with "wind," and "polar bears" are associated with "ice and snow." The majority of the "Inferences" items on CAT Level II, then, are well within the psycholinguistic processes of literal comprehension and therefore correlate more highly with the multiple-choice cloze than expected.

In summary, there is again a range of psycholinguistic processes subsumed under the label "Inferences" in CAT Level II, but an examination of the items reveals a preponderance of processes that fall within the construct, literal comprehension, hence the high correlation with the two experimental tests. (An examination of other disconfirmations in predicted correlational levels revealed a similar misleading application of labels to test items on the CAT.) CAT Levels I, II, III, and IV, therefore, appear in general to access more literal comprehension processes than the analysis of the subtest score labels would suggest. Inconsistencies in the expected pattern of intercorrelations, then, on closer examination reveal the consistency of the construct, literal comprehension, and the consistency of the behavior of the two experimental tests in spite of the misleading and inconsistent labels on the CAT subscores.

Further studies. The foregoing analysis is generally supportive of the accuracy and validity of the construct, literal comprehension, and its two experimental operationalizations, the multiple-choice cloze and the wh-item. But the foregoing analysis has also revealed inconsistencies in the application of labels to items in the CAT. Such inconsistencies are a necessary consequence of a test that is not theory-based. More meaningful analyses of the correlations among the literal comprehension tests and the CAT subscores, therefore, depend on an item-by-item analysis of the CAT, defining each item in relation to the explicit construct, literal comprehension, rather than attempting to interpret vague, global labels like "Facts" or "Inferences." Such an item-by-item analysis will also be a test of the explicitness and consequently the utility of the construct itself, that is, its ability to discriminate between items that appear to access different psycholinguistic processes and items that also behave differently in relation to the two operationalizations of the construct literal comprehension. Further studies, in other words, should lead to a refinement of both the construct and its operationalizations as well as the ability to identify what the tests actually measure.

Validity Across Alternate Measures of the Same Construct

The convergence of the two principal measures of literal comprehension as indicators of the same construct is evaluated in Table 17 which shows the parallel correlations of the multiple-choice cloze and wh-item tests with the various CAT scores by CAT level in the study sample. As before, confirmed predictions are underlined. Since the intent of the analysis is to examine the differences in the absolute values of the correlations of both measures of literal comprehension with relevant criteria, the correlations of these two measures with the tests of IQ have been included

at the bottom of the table.

Examination of the correlations in Table 9.17 reveals a pattern of remarkable consistency in the way the two very different indicators of the construct, literal comprehension, behave in relation to the range of psycholinguistic skills accessed by the CAT subscores. The differences between multiple-choice cloze and wh-item correlations are .05 or less for 34 out of 47 CAT subscores. In no case is the difference between the correlations greater than .09. The pattern of confirmation is thus also very consistent between the two tests of literal comprehension, there being only 5 out of 47 instances where there is lack of agreement on confirmation or disconfirmation. A similar pattern of consistency or convergence holds in the correlations of the wh-item and multiple-choice cloze tests with the measures of IQ.

The negligible differences in the way the two tests of literal comprehension compare in correlations across CAT levels and subscores is even more remarkable considering the differences in format and content between the two experimental tests. Besides radical differences in item type and format, the passages vary in content and length between the two experimental tests. Multiple-choice cloze passages are never more than 70 words long, while wh-item test passages contain as many as 220 words. Moreover, the content of the passages on the two tests is completely different. No passage that appears on the wh-item test appears anywhere on the multiple-choice cloze test. But both tests are designed to measure literal comprehension as defined in the construct regardless of variations in the subject matter, style, or length of the reading passages.

These data, taken together with the results of previous analyses

Table 17

Actual Zero-Order Correlations Among Literal Comprehension Test Scores and California Achievement Test Subscores by CAT Level (Language and Non-Language IQ Scores Included as Additional Criteria)

| | CAT I | | CAT II | | CAT III | | CAT IV | |
|---|---|---|---|---|---|---|---|---|
| | MCC[a] | WH[b] | MCC | WH | MCC | WH | MCC | WH |
| Vocabulary Subscores | | | | | | | | |
| Sentence-Pic. Assoc. | .06[c] | .06 | | | | | | |
| Beginning Sounds | .43 | .45 | | | | | | |
| Ending Sounds | .35 | .34 | | | | | | |
| Letter Recognition | .11 | .13 | | | | | | |
| Word Form | .27 | .34 | | | | | | |
| Picture-Word Assoc. | .53 | .60 | | | | | | |
| Word Recognition | .37 | .45 | .56 | .61 | | | | |
| Words in Context | .65 | .70 | .76 | .78 | .70 | .64 | .57 | .50 |
| Comprehension Subscores | | | | | | | | |
| Facts | .60 | .64 | .76 | .75 | .67 | .66 | .45 | .43 |
| Interpretation | .52 | .61 | .73 | .69 | .70 | .64 | .50 | .50 |
| Relationships | | | | | .43 | .44 | .38 | .42 |
| Generalizations | | | .73 | .67 | .51 | .47 | .42 | .41 |
| Inferences | .48 | .53 | .73 | .70 | .35 | .32 | .44 | .44 |
| Reading-General | | | | | .62 | .57 | .46 | .46 |
| Reading-Soc. Studies | | | | | .58 | .59 | .46 | .48 |
| Reading-Science | | | | | .58 | .56 | .42 | .43 |
| Reading-Math. | | | | | .51 | .44 | .40 | .37 |
| Language Subscores | | | | | | | | |
| Standard English | .44 | .52 | .73 | .69 | .46 | .42 | .31 | .29 |
| Sentence Structure | | | | | .37 | .32 | .42 | .37 |
| Sent. Parts and Funct. | | | | | .31 | .23 | .45 | .37 |
| Transformation | | | | | .28 | .23 | .22 | .32 |
| Language IQ | .34 | .41 | .55 | .53 | .62 | .56 | | |
| Non-Language IQ | .43 | .47 | .49 | .46 | .52 | .50 | | |

[a] Multiple-choice cloze test.

[b] Wh-item test.

[c] Underlined values were within level of correlation predicted.

reported here of the correlations among subscores and total scores on the multiple-choice cloze and wh-item tests, provide rather strong confirmation of the validity and generalizability of the trait in question.

Results of the Factor Analyses

This final section of this exploratory analysis provides additional data on the proposition that literal comprehension as measured by the cloze and wh-item tests may be discriminated from other forms of reading comprehension, namely the inferential and related reasoning processes that seem to be emphasized in standardized tests of reading comprehension. The present analysis also attempts to corraborate the distinctions made previously about the psycholinguistic content of the comprehension section of the CAT, thus lending further credence to the analysis in the preceding section. That is, the CAT seems to include both literal and inferential types of items under the same rubric of inferences in the comprehension section of the test.

The distinction between literal and inferential processes as related but clearly different types of comprehension was made by Carroll in a conceptual paper on types of comprehension. Carroll considered:

> . . . whether it is possible to in fact to distinguish "pure" comprehension of language texts from processes of inference, deduction, and problem solving that often accompany the reception of language. An empirical research question would be to see whether it would be possible to decrease the correlation of comprehension ability with intelligence tests by eliminating or reducing those elements of comprehension that call for inferential processes that go beyond sheer comprehension (1972, p. 37).

The present study affords a unique opportunity of testing Carroll's hypothesis through exploratory factor analysis, since prior attempts to factor comprehension into separate abilities have not adequately represented the literal component in the form of appropriate and sufficient numbers of items or have failed to combine the literal component in the analysis with

standardized tests of reading comprehension (Weaver and Kingston, 1963; Bormuth, 1969; Ohnmacht, Weaver, and Kohler, 1970; Spearritt, 1972; Thorndike, 1973-74; Davis, 1974). The multiple-choice cloze and wh-item tests available here, along with appropriate subtests from the CAT, would appear to be an adequate basis for the formation of a literal comprehension factor. Similarly, an adequate basis seems to exist for defining an inferential factor in appropriate subtests of certain levels of the CAT and the IQ measures. However, since the inferential CAT subtests seem to be heavily contaminated with literal comprehension, these subtests might be expected to "split" between the hypothesized literal and inferential comprehension factors. Additional factors may be resolved around the language sections of the CAT and around the Level I sections relating to pre-reading skills.

Analysis. The measures included in the factor analysis were the 4 grammatical deletion-type subscores for the multiple-choice cloze, the 8 wh-subscores for the wh-item test, the language and non-language IQ scores, the CAT reading subscores listed previously in Table 6, and the CAT language subscores and mechanics skill score. This procedure provided at least two test scores to mark each potential factor, including the possibility that the cloze and wh-item tests might each form separate factors.

The data from the foregoing tests were intercorrelated by subsample and the resulting matrices subjected to principal components analyses with ones in the diagonal. Components with Eigenvalues greater than 1.00 were then rotated to the varimax criterion. This procedure was followed for each of the four subsamples with the exception that, at Levels I and II, the data were rerun with only the noun and verb scores used to represent the cloze test in order to obtain a clearer resolution of the factor structure.

Results. Tables 18-22 present by subsample the factor solutions rotated to the varimax criterion. The factor structure for the Level I subsample is, as noted, shown twice: (a) first in Table 18 with the cloze noun and verb scores to mark the multiple-choice cloze test; and (b) secondly in Table 19 with all four deletion-type or grammatical subscores in the cloze test. The additional factor matrix in Table 19 is for the reader's convenience and is intended to illustrate the tendency for the cloze test to generate two factors in the Levels I and II subsamples. One factor is the expected literal comprehension factor and is marked by the noun and verb subscores along with other appropriate test scores. The second factor is a relatively unimportant cloze factor with the highest loadings for the adjective and adverb score. This second factor may reflect language comprehension processes different from what is reflected by the noun-verb score, or it could represent developmental differences in what is measured by the multiple-choice cloze in the lower grades. Since the appearance of this additional factor resulted in problems in interpreting the results of the factors analyses at Levels I and II, only the analyses using the noun and verb scores to represent the cloze test are interpreted for these levels. In the upper grade samples, all four of the deletion type subscores are used to represent the multiple-choice cloze test.

Turning now to the factor solution in Table 18, the results for grade 1 show 3 rotated factors. The first factor, which is interpreted as literal comprehension, is defined in order of importance by the CAT Words in Context subtests, the cloze noun score, the CAT comprehension subtests, the cloze verb score, the various wh-item scores, and the CAT subtests for Picture-Word Association, Language Mechanics, and Language Usage. The last three subtests are not of particular importance in defining the factor. As

TABLE 18

Varimax Rotated Loadings for Multiple-Choice Cloze, Wh-item, IQ, and CAT Skill and Subscores: Level I (N = 456)

| Subscores | Factor I | Factor II | Factor III | Final Communality |
|---|---|---|---|---|
| **Multiple-Choice Cloze:** | | | | |
| Noun | .73 | -.24 | .06 | .59 |
| Verb | .64 | -.27 | .05 | .48 |
| **Wh-item:** | | | | |
| How | .65 | -.08 | -.04 | .43 |
| What (N) | .63 | -.26 | .02 | .47 |
| What (V) | .63 | -.20 | -.07 | .43 |
| When | .53 | -.10 | .18 | .32 |
| Where | .62 | -.12 | .01 | .40 |
| Which | .66 | -.19 | .01 | .48 |
| Who | .65 | -.26 | .04 | .49 |
| Why | .52 | -.15 | .06 | .30 |
| **IQ:** | | | | |
| Language | .28 | -.47 | .45 | .50 |
| Non-Language | .37 | -.50 | .32 | .49 |
| **Vocabulary:** | | | | |
| Sentence-Picture Assn. | -.06 | -.08 | .83 | .71 |
| Beginning Sounds | .30 | -.77 | .09 | .69 |
| Ending Sounds | .17 | -.73 | .03 | .56 |
| Letter Recognition | -.06 | -.64 | -.17 | .44 |
| Word Forms | .16 | -.67 | .18 | .51 |
| Picture-Word Assn. | .48 | -.69 | .05 | .71 |
| Word Recognition | .24 | -.79 | .08 | .68 |
| Words in Context | .77 | -.29 | .15 | .70 |
| **Comprehension:** | | | | |
| Facts | .72 | -.24 | .14 | .60 |
| Interpretation | .70 | -.21 | .06 | .54 |
| Inferences | .69 | -.01 | .18 | .51 |
| Language Mechanics | .45 | -.51 | .33 | .57 |
| **Language Usage:** | | | | |
| Standard English | .39 | -.64 | .29 | .64 |
| Eigenvalue | 9.85 | 2.34 | 1.05 | |
| Cumulative Proporation of Total Variance | .39 | .49 | .53 | |

Table 19

Varimax Rotated Loadings for Multiple-Choice Cloze, Wh-item, IQ, and CAT Skill and Subscores: Level I (N = 456)

| Subscores | Factor I | II | III | IV | Final Communality |
|---|---|---|---|---|---|
| Multiple-Choice Cloze: | | | | | |
| Noun | -.58 | .25 | -.54 | .03 | .69 |
| Verb | -.48 | .28 | -.55 | .03 | .61 |
| Adjective | -.21 | .07 | -.75 | .06 | .61 |
| Adverb | -.07 | .02 | -.71 | -.01 | .51 |
| Wh-item: | | | | | |
| How | -.67 | .03 | -.18 | -.05 | .48 |
| What (N) | -.64 | .27 | -.10 | .01 | .49 |
| What (V) | -.65 | .20 | -.04 | -.08 | .47 |
| When | -.54 | .10 | -.06 | .17 | .34 |
| Where | -.65 | .12 | -.02 | .01 | .44 |
| Which | -.66 | .19 | -.13 | .00 | .49 |
| Who | -.69 | .26 | .01 | .04 | .54 |
| Why | -.55 | .13 | -.07 | .05 | .32 |
| IQ: | | | | | |
| Language | -.29 | .48 | -.02 | .44 | .51 |
| Non-Language | -.36 | .51 | -.09 | .30 | .49 |
| Vocabulary: | | | | | |
| Sentence-Picture Assn. | .06 | .10 | -.03 | .84 | .72 |
| Beginning Sounds | -.25 | .78 | -.17 | .07 | .70 |
| Ending Sounds | -.11 | .73 | -.17 | .02 | .58 |
| Letter Recognition | .06 | .63 | .05 | -.18 | .44 |
| Word Forms | -.16 | .67 | -.02 | .17 | .51 |
| Picture-Word Assn. | -.45 | .70 | -.14 | .04 | .71 |
| Word Recognition | -.23 | .79 | -.06 | .07 | .68 |
| Words in Context | -.70 | .31 | -.31 | .13 | .69 |
| Comprehension: | | | | | |
| Facts | -.64 | .26 | -.35 | .12 | .61 |
| Interpretation | -.66 | .22 | -.23 | .04 | .54 |
| Inferences | -.63 | .03 | -.27 | .17 | .50 |
| Language Mechanics | -.42 | .52 | -.16 | .31 | .57 |
| Language Usage: | | | | | |
| Standard English | -.39 | .65 | -.05 | .27 | .65 |
| Eigenvalue | 10.08 | 2.42 | 1.27 | 1.05 | |
| Cumulative Proporation of Total Variance | .37 | .46 | .51 | .55 | |

expected, the IQ subscores load at very low levels with factor I.

The second factor is composed largely of phonological, orthographic and word recognition skills, with moderate loadings on the factor for the IQ subscores and the language test scores. This factor seems to reflect a combination of the pre-reading skills and general verbal ability that are important components in learning to read.

The third factor is primarily defined by the CAT Sentence-Picture Association subtest. The pattern of results for Level I is considered to be generally supportive of theoretical expectations although there is the apparent inconsistency of the loading of the CAT Inferences subtest on Factor I and the failure of an "inferences" factor to appear in the matrix. These inconsistencies seem to be resolved by the fact that inferential processes are represented only weakly in the Level I CAT comprehension section, there being only 8 items thus classified. Moreover, as argued in the preceding section, many of these items are doubtful measures of inferential processes.

The rotated factor matrix for Level II shown in Table 20 provides more definitive support for theoretical expectations. The first factor seems to be clearly a literal comprehension factor with highest loadings on the cloze and wh-item tests, and the CAT Word Recognition, Words in Context, and Facts subtests. The IQ subtests are virtually uncorrelated with Factor I and the inferential subtests have moderate loadings with the factor. In contrast, factor II, which appears to define an inferential factor, shows higher loadings on the CAT subtests with inferential items. The notion that the inferential subtests confound both literal and inferential items is thus supported by the clear tendency for these subtests to "split" between factors I and II.

TABLE 20

Varimax Rotated Loadings for Multiple-Choice Cloze, Wh-item, IQ, and CAT Skill and Subscores: Level II (N = 972)

| Subscores | Factor I | Factor II | Final Communality |
|---|---|---|---|
| Multiple-Choice Cloze: | | | |
| Noun | .61 | .60 | .72 |
| Verb | .60 | .58 | .70 |
| Wh-item: | | | |
| How | .70 | .20 | .52 |
| What (N) | .65 | .28 | .50 |
| What (V) | .69 | .22 | .52 |
| When | .56 | .23 | .37 |
| Where | .67 | .22 | .49 |
| Which | .73 | .20 | .57 |
| Who | .73 | .22 | .59 |
| Why | .61 | .29 | .46 |
| IQ: | | | |
| Language | .16 | .78 | .63 |
| Non-Language | .06 | .79 | .62 |
| Vocabulary: | | | |
| Word Recognition | .61 | .35 | .49 |
| Words in Context | .70 | .54 | .78 |
| Comprehension: | | | |
| Facts | .65 | .61 | .79 |
| Interpretation | .56 | .64 | .73 |
| Generalizations | .53 | .65 | .71 |
| Inferences | .59 | .61 | .72 |
| Language Mechanics: | .45 | .65 | .62 |
| Language Usage: | | | |
| Standard English | .58 | .62 | .72 |
| Eigenvalue | 11.07 | 1.21 | |
| Cumulative Proportion of Total Variance | .55 | .61 | |

It should also be noted that Factor II appears to have a fairly strong verbal ability component as evidenced by the high loading for language IQ and the moderately high loadings for the CAT language skill and subtest scores. The fact that the cloze scores also have moderate loadings on Factor II suggests that there is a minor verbal ability component in the multiple-choice cloze test.

Table 21 presents the rotated factor matrix for the Level III subsample. The results here generally parallel those obtained for Level II, with the additional qualifications that literal comprehension, the first factor, is more strongly defined by the cloze and wh-item subscores, and the language factor noted in the Level II analysis appears here as a separate factor. The inferential components of the CAT now appear to be divided among all three factors, but load most strongly on factor III, which is clearly the inference factor. The IQ subscores load very highly on factor III, and have low loadings on factors I and II.

Finally, Table 22 presents the rotated factor matrix for the CAT Level IV subsample. Four factors appear that are identifiable as CAT comprehension, comprehension as measured by the multiple-choice cloze, comprehension as measured by the wh-item test, and a language factor composed of the CAT language subtests. This factor pattern is generally inconsistent with the results for the other three levels. It is difficult to explain this pattern with any assurance, but the interpretation most suggested by the data is that the expected psycholinguistic resolution of the factors was overwhelmed by the different testing methodologies used in the study. Contributing to this hypothetical resolution of the analysis in the form of "method" factors was the lack of an IQ test score at Level IV and generally lower positive correlations among the variables in the analysis.

Table 21

Varimax Rotated Loadings for Multiple-Choice Cloze, Wh-item, IQ, and CAT Skill and Subscores: Level III
(N = 1697)

| Subscores | Factor I | Factor II | Factor III | Final Communality |
|---|---|---|---|---|
| Multiple-Choice Cloze: | | | | |
| Noun | .71 | .35 | .33 | .73 |
| Verb | .70 | .35 | .32 | .72 |
| Adjective | .64 | .33 | .29 | .61 |
| Adverb | .56 | .25 | .25 | .44 |
| Wh-item: | | | | |
| How | .70 | .10 | .13 | .52 |
| What (N) | .64 | .18 | .11 | .46 |
| What (V) | .61 | .10 | .24 | .44 |
| When | .65 | .09 | .17 | .45 |
| Where | .67 | .09 | .14 | .48 |
| Which | .71 | .11 | .17 | .54 |
| Who | .68 | .09 | .16 | .49 |
| Why | .67 | .16 | .17 | .50 |
| IQ: | | | | |
| Language | .27 | .09 | .82 | .76 |
| Non-Language | .23 | .06 | .81 | .72 |
| Vocabulary: | | | | |
| Words in Context | .47 | .42 | .59 | .75 |
| Comprehension: | | | | |
| Facts | .56 | .30 | .49 | .64 |
| Interpretation | .51 | .42 | .52 | .71 |
| Relationships | .27 | .32 | .47 | .40 |
| Generalizations | .28 | .40 | .56 | .56 |
| Inferences | .11 | .29 | .54 | .39 |
| Language Mechanics | .36 | .65 | .35 | .68 |
| Language Usage: | | | | |
| Standard English | .28 | .62 | .21 | .50 |
| Sentence Structure | .19 | .67 | .07 | .49 |
| Sentence Parts and Functions | .04 | .64 | .19 | .45 |
| Transformation | .06 | .65 | .09 | .44 |
| Eigenvalue | 10.82 | 1.84 | 1.18 | |
| Cumulative Proportion of Total Variance | .43 | .51 | .55 | |

Table 22

Varimax Rotated Loadings for Multiple-Choice Cloze, Wh-Item, and CAT Skill and Subscores: Level IV (N = 1260)

| Subscores | I | II | III | IV | Final Communality |
|---|---|---|---|---|---|
| Multiple-Choice Cloze: | | | | | |
| Noun | -.27 | .35 | -.80 | .14 | .85 |
| Verb | -.24 | .35 | -.78 | .18 | .81 |
| Adjective | -.28 | .26 | -.79 | .13 | .79 |
| Adverb | -.14 | .24 | -.73 | .11 | .62 |
| Wh-item: | | | | | |
| How | -.12 | .71 | -.19 | .10 | .55 |
| What (N) | -.13 | .72 | -.11 | .09 | .55 |
| What (V) | -.19 | .66 | -.14 | .16 | .52 |
| When | -.16 | .69 | -.19 | .12 | .55 |
| Where | -.21 | .63 | -.22 | .10 | .50 |
| Which | -.23 | .68 | -.13 | .12 | .54 |
| Who | -.17 | .73 | -.21 | .08 | .61 |
| Why | -.24 | .67 | -.21 | .08 | .56 |
| Vocabulary: | | | | | |
| Words in Context | -.74 | .24 | -.30 | .20 | .73 |
| Comprehension: | | | | | |
| Facts | -.66 | .23 | -.20 | .09 | .54 |
| Interpretation | -.76 | .26 | -.18 | .21 | .72 |
| Relationships | -.73 | .23 | -.08 | .09 | .60 |
| Generalizations | -.71 | .16 | -.16 | .20 | .60 |
| Inferences | -.80 | .21 | -.13 | .16 | .72 |
| Language Mechanics | -.35 | .21 | -.25 | .74 | .77 |
| Language Usage: | | | | | |
| Standard English | -.21 | .11 | -.09 | .77 | .66 |
| Sentence Structure | -.41 | .14 | -.24 | .47 | .47 |
| Sentence Parts and Functions | -.39 | .12 | -.31 | .47 | .48 |
| Transformations | -.01 | .13 | -.01 | .75 | .57 |
| Eigenvalue | 9.53 | 2.09 | 1.37 | 1.32 | |
| Cumulative Proportion of Total Variance | .41 | .51 | .56 | .62 | |

## Summary and Conclusions

The test development project reported here has evolved through a number of years of research and development on major issues relating to the assessment of school achievement. The intent of this project is to produce a test development resource that will be useful at a variety of instructional levels and a measure that will be unique in at least two major respects: (1) It will be a measure of reading comprehension that is referenced to domains of written discourse, and (2) It will be the only extant and broadly applicable measure of literal comprehension as such--the important, generalized reading skill that is hypothesized to underlie all higher-order reading comprehension abilities. Some two years of development effort have culminated in large pools of reading comprehension items and passages that are referenced to several major domains of reading materials relevant to students in grades 1-12. As currently conceived, these item pools constitute a basis for assembly of large numbers of tests of literal comprehension in grades 1-12, across several major content domains relevant to the school population.

The content of the item pool developed to this point consists of the multiple-choice cloze component and an alternate measure of the construct of literal comprehension based on the wh-item. The multiple-choice cloze component consists of approximately 1,500 clozed passages (generally, 60-70 word passages with ten deletions and accompanying multiple-choice items) categorized (temporarily) by readability levels determined by Spache and Dale-Chall readability formulas. The wh-/main idea Passages vary systematically in length by readability level (e.g., approximately 25 words at level 1 and up to 220 words at levels 17-20). Each of these passages is accompanied by as many as four multiple-choice main idea items

and up to eight multiple-choice wh-detail items modeled after Bormuth's (1970) wh-items. The formats of the cloze and wh-materials are both objective, generative procedures for preparing numbers of parallel, multiple-choice items.

The first field test of the cloze and wh-item tests was conducted in May 1975, in an administration of both types of tests in a survey design to approximately 5,000 students spread more or less evenly over grades 1-9. This administration fulfilled several purposes: (1) It explored the use of the testing materials in applying a survey design in one textual area; (2) it provided data for detailed item analyses; (3) it provided an initial test of the ability of the system in assembling large numbers of parallel test forms; (4) it made available reliability data on a large number of test forms; and (5) it provided initial evidence on the validity of the construct underlying the cloze and wh-item tests. The more important conclusions that were drawn from the field test are as follows:

1. The existing paper-based model of the item pool allowed assembly of large numbers of parallel test forms in a matter of a few hours.

2. The application of the survey design model in grades 1-9 was generally successful, for both the cloze and wh-item tests, but the design can be improved in the future by raising the ceiling of readability for upper-grade test batteries.

3. Large numbers of virtually parallel tests could be systematically assembled from either the cloze or wh-item passage and item pools. With improved scaling, the possibility of objectively assembling n tests with specified properties is assured, thus providing for transferability of test generation.

4. Analyses of the reliability of the cloze test forms showed that the tests assembled for the study were highly precise across all grade levels in the study sample. The level of precision is sufficiently high to warrant use of the tests at the individual level. The reliability data further support the infer-

ence that the cloze test is reliable over short intervals (i.e., alternate forms of the same test will scale individuals similarly on test-retest with a high degree of precision). The reliability characteristics of the wh-item tests were similar to those achieved with the cloze test.

5. There were several indications of support of the construct validity of the cloze test in the correlational analyses. The internal consistency measures indicated the cloze test could be accurately described as measuring a homogeneous trait across grades 1-9. The validity coefficients between the cloze test and the wh-item test, an alternate measure of the construct, were consistently high ($r$ = .81 at grades 1-3), except where attenuated by range of talent. The cloze test generally correlated at appropriate levels with measures of verbal and non-verbal IQ, California Achievement Test (CAT) subscores in language and reading, and a measure of passage dependency. The cloze and wh-item tests converged in having virtually identical correlations with the CAT subscores, the IQ scores, and the score on passage dependency. Overall, the results were highly consistent across the 9 grade levels, lending considerable credibility to the validity of the cloze test.

6. The results of a series of factor analyses run on the test score matrix for four levels of the sample were generally supportive of the reasoning underlying the test. With the exception of the level IV results, the factor analyses generated a first factor that was clearly interpretable as literal comprehension and that was defined by the multiple-choice cloze, the wh-item test, and appropriate CAT subtests. A second factor, marked by high loadings on IQ and moderate loadings on inferential subtests of the CAT, was identifiable as an inferential reasoning factor.

Considering that this is the exploratory phase of a research project on new approaches to testing comprehension, the results are very encouraging in several respects. It appears that the standard cloze technique has here been successfully converted into a measure of reading comprehension that preserves the original advantages of the cloze methodology while offering the additional advantage of the multiple-choice format. Data has also been provided that suggest the multiple-choice cloze is potentially a more valid measure of language comprehension *per se* as distinguished from reading as a process of reasoning (Carroll, 1972). Considerable evidence has been compiled to show that the standard cloze format emphasizes the syntactic

component of a test passage at the expense of the semantic component and that the test is substantially an indirect measure of higher order reasoning competencies (Rankin, 1974). It is no surprise then that the standard cloze correlates substantially with standardized measures of reading comprehension, but as should be evident from the attempt to use the CAT as a criterion for comprehension here, such relationships may not constitute an adequate basis for claims about the validity of the standard cloze as a measure of language comprehension.

Further study of the issue of the validity of the multiple-choice cloze and other alternate formats of the standard cloze is much needed. An array of additional comprehension and language criteria need to be constructed to illuminate more precisely the psycholinguistic meaning of various cloze item formats (e.g., paraphrase transformations). In addition, it will be necessary to define more carefully the variation in semantic and syntactic factors that are sampled in the cloze passages assembled into a test for a student. There are also a number of unresolved developmental issues relating to the new cloze test format. For example, if the new multiple-choice cloze format is a measure of language comprehension per se, the test should then be highly responsive to the language and reading experiences of children in and out of school. Obviously, much work remains before the cloze becomes a generally useful methodology for testing reading comprehension, but the obvious advantages of the technique from both theoretical and applied points of view indicate the desirability of continuing the task.

REFERENCES


Bormuth J. Development of standards of readability: Toward a rational criterion of passage performance. Final Report (U.S. Office of Education: Project No. 9-0237). Chicago, Ill.: University of Chicago, 1971, (ERIC Document Reproduction Service No. ED 054-233).

Bormuth, J. Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly, 1969, 4, 358-365.

Bormuth, J. On the theory of achievement test items. With an appendix by P. Menzel: On the linguistic bases of the theory of writing items. Chicago: University of Chicago Press, 1970.

California Achievement Test: Test coordinator's handbook. Monterey, Calif.: CTB/McGraw-Hill, Del Monte Research Park, 1970.

Carroll, J. Defining language comprehension: Some speculations. In R. Freedle & J. Carroll (Eds.), Language comprehension and the acquisition of language. New York: Wiley, 1972.

Carroll, J., Davies, P., & Richman, B. The American Heritage word frequency book. New York: Houghton-Mifflin, 1971.

Cranney, A. The construction of two types of cloze reading tests for college students. Journal of Reading Behavior, 1972, 5, 60-64.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrics, 1951, 16, 297-334.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Dale, E., & Chall, J. A formula for predicting readability: Instructions. Educational Research Bulletin, 1948, 27, 37-47.

Davis, F. B. Fundamental factors of reading among elementary-school pupils: A preliminary report. In P. L. Nacke (Ed.), Interaction: Research and practice for college-adult reading, Twenty-third Yearbook of the National Reading Conference, Clemson, S.C.: National Reading Conference, 1974.

Fillenbaum, S., Jones, L., & Rapoport, A. The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 186-194.

Hansen, L.H., & Hesse, K. D. A pilot reading literacy assessment of Madison Public School students: Final report. Madison, Wisconsin: The Madison, Wisconsin Public Schools, The Department of Research and Development, 1974.

Harris, A., & Jacobson, M. Basic elementary reading vocabularies. New York: Macmillan, 1972.

Hively, W. Domain-referenced testing: Part one, basic ideas, introduction to domain-referenced testing. Educational Technology, 1974, 14 (6), 5-10

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Katz, J. Semantic theory. New York: Harper & Row, 1972.

Katz, J. & Fodor, J. The structure of a semantic theory. Language, 1963, 39, 170-210.

Katz, J., & Fodor, J. The structure of a semantic theory. In J.P. DeCecco (Ed.), The psychology of language, thought, and instruction: Readings. New York: Holt, Rinehart, & Winston, 1967.

Kuder, G.F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Ohnmacht, F., Weaver, W., & Kohler, E. Cloze and closure: A factorial study. The Journal of Psychology, 1970, 74, 205-217.

O'Reilly, R. P., Kidder, S. J., & Schuder, R. T., A Proposal for Validation and Refinement of Measures of Literal Comprehension in Reading for Use in Policy Research and Classroom Management (Photocopy). Albany, N.Y.: The University of the State of New York, The State Education Department, Division of Research, 1976.

Rankin, E. The cloze procedure revisited. In P.L. Nacke (Ed.), Interaction: Research and practice for college-adult reading, Twenty-third Yearbook of the National Reading Conference, Clemson, S.C.: National Reading Conference, 1974.

Smith, F. Comprehension and learning: A conceptual framework for teachers. New York: Holt, Rinehart, & Winston, 1975.

Spache, G. A new readability formula for primary-grade reading materials. Elementary School Journal, 1953, 53, 410-413.

Spache, G. Good reading for poor readers. Champaign, Ill.: Garrard Press, 1960.

Spearritt, D. Identification of subskills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly, 1972, 8, 92-111.

Taylor, S., Frackenpohl, H., & White, C. A revised core vocabulary: A basic vocabulary for grades 1-8, an advanced vocabulary for grades 9-13 (Research and Information Bulletin No. 5). Huntington, N.Y.: Educational Developmental Laboratories, March 1969.

Thorndike, R. Reading as reasoning. Reading Research Quarterly, 1973-1974 9, 135-147.

Tuinman, J. Determining the passage dependency of comprehension questions in 5 major tests. Reading Research Quarterly, 1973-1974, 9, 206-223.

Weaver, W., & Kingston, A. A factor analysis of the cloze procedure and other measures of reading and language ability. Journal of Communication, 1963, 13, 252-261.